ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTITIONAL SCIENCESCHOOL OF INFORMATION SCIENCE

APPLYING DATA MINING TECHNIQUES FOR CUSTOMERS SEGMENTATION AND PREDICTION: THE CASE OF DASHEN BANK

By

Etsegenet Gebregiorgies ID: GSE8325/13

Advisor: Million Meshesha (PhD)

Submitted to the School of Information Sciences in PartialFulfillment for the Degree of Master of Science in Information Science and systems

June, 2023

ADDIS ABABA UNIVERSITY

COLLEGE OF NATURAL AND COMPUTITIONAL SCIENCESCHOOL OF INFORMATION SCIENCE

APPLYING DATA MINING TECHNIQUES

FOR CUSTOMERS SEGMENTATION AND PREDICTION

BY

ETSEGENET GEBREGIORGIES

Approved by Examination Board

_____	_____	_____
Chairman		
_____	_____	_____
Supervisor		
_____	_____	_____
Internal Examiner		
_____	_____	_____
External Examiner		

June, 2023

DECLARATION

I affirm that the research titled "Applying Data Mining Techniques for Customers Segmentation and Prediction " is my own work, done under the guidance of Dr. Million Meshesha. This research has not been previously presented in any forum, and all sources utilized have been duly acknowledged.

Etsegenet Gebregiorgies _____ _____

I certify that the above statements made by the student are correct.

_____ _____ _____

Supervisor Signature Date

ACKNOWLEDGEMENTS

I want to start by thanking the Almighty God for making it possible for me to advance in my academic studies. I would not have progressed this far without His favor. Second, I want to express my sincere gratitude to Dr. Million Meshesha, my thesis advisor, for his invaluable assistance and unwavering support throughout my research. His advice and support were extremely helpful to me as I finished my thesis.

Dr. Million Meshesha diligent supervision, feedback, and constructive criticism on my work helped me to stay focused, motivated, and confident. His insightful comments and guidance aimed at directing my research work towards the right direction and ensuring that it met the required academic standards.

I would also like to express my gratitude to Ato Dawit Sernessa, Senior Manager of Analytics Department at Dashen Bank, for his kind advice and support during the data collection and domain expert consultation. Furthermore, I am grateful to all the staff members of the Data Analytics department at Dashen Bank for their support and feedback on data collection and attribute selection.

I extend a special thanks to Ato WeldeMariam Dresses, Director of Strategic Research & Corporate Performance Department, for his continuous guidance and support on business strategy and data collection.

Lastly, I want to express my heartfelt appreciation to my family for their emotional support and encouragement towards my academic journey. Their unwavering support has motivated me to work hard and achieve my dream.

ABSTRACT

The banking sector has changed significantly in how it does business, putting a greater emphasis on contemporary technology to stay competitive. The banking sector has begun to understand how critical it is to build a knowledge base and use it to the bank's advantage in the field of strategic planning. Finding clients who are more likely to be interested in a product or service is a crucial task. Data mining has been widely used for customer segmentation and identification in order to predict potential customers for a given product or service.

This study uses a six-step hybrid Knowledge Discovery Process model with the goal of applying data mining for the purpose of customer segmentation and prediction. The necessary data was gathered from the Bank's CBS database, and then pre-processing operations like data transformation and cleansing were performed in order to produce high-quality data for use in data mining with WEKA software. The goal of this thesis is to create a model that can be used to categorize Dashen Bank customers based on their transactional data and forecast which customers will be profitable for the bank. Since there are no predefined classes that describe the customers of the bank, the researcher uses clustering techniques (such as K-means, Filtered cluster and Farthest First) that resulted in the appropriate number of clusters for customer segmentation. K-means clustering, which divides potential customers based on their monthly credit turnover, produces the best descriptive model. By labeling the unlabeled data set as a result of clustering, classification algorithms like J48 Decision Trees, K Nearest Neighbor (KNN), and Naive Bayes can be used to build a model that allows for customer prediction. Researchers divide the data into three distinct clusters based on the transactional amount range by using a clustering algorithm. The labels "SMALL," "MEDIUM," and "CORPORATE" are applied to these clusters. "The range of transaction is the main distinction between these clusters.

Experimental result shows that, out of the three algorithms, J48 decision tree with 70/30 test mode have the highest performance accuracy of 92.08%, which is selected for customer prediction. After consulting with experts, the data mining analysis revealed intriguing and unexpected attributes and patterns. The findings indicate that customers who exhibit basic deposit behavior with an overdraft facility are classified as corporate transactional customers. On the other hand, customers who hold a USD account are also categorized as corporate customers, which results in higher profitability for the bank. The study is based on only transactional data and customers are segmented on their monthly activities. To get 360 customers view, further research needs o be done towards coming up with customer profiling and customer relationship management (CRM) system

Keywords: Dashen Bank; Data Mining; Customer Segmentation; Customer Prediction

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ACRONYMS AND ABBREVIATION

ARFF	Attribute-Relation File Format
CBS	Core Banking System
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSV	Comma Separated Value
IFB	Islamic Finance Bank
KDD	Knowledge Discovery in Databases
NBE	National Bank of Ethiopia
PL/SQL	Procedural Language/Structured Query Language,
SSE	Sum Squared Error

WEKA Waikato Environment for Knowledge Analysis

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

The fundamental idea of business and how business operations are conducted have been completely altered by the computerization of financial operations, the use of the internet, and automated software. There is no exception to this in the banking industry. The way banking operations are conducted has also undergone a significant change [1]. Electronic points of contact are taking the place of conventional face-to-face customer interactions to speed up application processing, lower costs, and ultimately improve financial performance [2]. Today, data might be one of a bank's most valuable resources—but only if the bank knows how to uncover and make use of the valuable information that lies hidden in the raw data. Data mining enables knowledge extraction from historical data and future scenario prediction [3]. It aids in decision-making that is optimized for business, raises the importance of every client and communication, and boosts client satisfaction.

The process of extracting useful hidden knowledge from enormous amounts of data is known as data mining [3]. Data analysis, knowledge extraction, and knowledge mining from databases are just a few of the other terms that are used to describe data mining [3]. Nowadays, it is generally acknowledged that data mining is a critical step in the knowledge discovery in databases (or KDD) process. Three types of tasks are frequently involved in data mining [3]. The first is classification, which divides the data into predetermined categories and builds a mapping function between input and output. The other is clustering, in which the groups aren't predefined and the algorithm tries to group things that are similar together. In addition, association rule learning looks for undiscovered connections between variables [4].

Many businesses today use data mining applications for a variety of purposes, and one of these applications is customer segmentation, which enables businesses to pinpoint the precise characteristics of their product and service buyers and develop effective business strategies. Customer segmentation is now a valuable tool in the banking industry for both acquiring new customers and increasing the value of those already on board. Customer segmentation is the process of grouping customers according to a variety of attributes and traits into distinct, significant, and homogeneous subgroups [5]. It serves as a marketing tool for differentiation. It enables businesses to comprehend their clients and develop distinctive strategies [5]. In order to help businesses, become more practical and knowledge-driven, data mining techniques are also used to forecast future probabilities and behaviors [6]. Numerous studies have supported the value of customer segmentation [7]. Senior bank managers' rankings of the current retail banking priorities for their institutions in a global survey of retail banking trends are one example of such confirmation [7]. These findings indicate that improving customer segmentation and taking it into account when

designing and distributing new products is one of the top priorities (and a prerequisite for a successful future) [7].

According to Ziafat, Majid and Shakeri [5], there are various criteria for customer segmentation that can be used to optimize consumer marketing. The most widely used customer segmentation types are the following [5].

1. **Value based:** Customers are categorized in value-based segmentation based on factors like potential revenue and disposable income. Since it can be used to pinpoint the most valuable clients and track changes in value over time, this is one of the most crucial segmentation types.
2. **Behavioral**: This kind of segmentation is practical and highly effective. Also, it is extensively utilized since it poses few challenges in terms of data accessibility, including their purchasing behaviors, brand interactions, and patterns of product usage.
3. **Propensity based**: In propensity-based segmentation, customers are categorized based on propensity scores, such as cross-selling and churn scores, which are estimated by the appropriate classification (propensity) models.
4. **Socio-demographic and life-stage**: Based on sociodemographic and/or life-stage data, such as gender, race, age, social status, and education, this type shows various customer groupings.
5. **Needs/attitudinal**: Based on customer needs, wants, attitudes, preferences, and perceptions of the company's services and products, this type of segmentation categorizes customers into different groups. Typically, it is based on data from market research.

In contrast, organizations need a comprehensive understanding of their customers to have a competitive advantage. The needs, wants, attitudes, behaviors, preferences, and perceptions of their clients must also be a priority. For this, it is necessary to analyze pertinent data in order to pinpoint the underlying customer segments. With the help of customized product offerings and marketing campaigns, among other things, the organization will be able to manage and target groups with particular characteristics more successfully. Data mining applications that cater to seasoned clients include customer segmentation [5]. Defining business goals is the first step in customer segmentation, which is followed by delivering tailored marketing strategies to each segment. Based on the specific criterion or attributes used for segmentation, there are numerous different segmentation types. Customers can be divided into groups based on the value they provide. Depending on the specific business goal, a particular segmentation technique is used [8].

Segmentation analysis is rarely restricted to just one or a few variable by smart marketers. Instead, they employ various segmentation bases in an effort to pinpoint more narrowly defined target groups. As a result, a bank may not only recognize a group of wealthy clients but may also look for several distinguishing characteristics within that group, such as income, assets, saving, risk preferences, housing, and lifestyles. Even more, it's important to track and carefully observe changes in segmentation variables like income, age, and motivation before deciding whether to keep or eliminate a particular segment.

The selection of data mining as a methodology for this study is based on the researcher's expertise and knowledge in this field. Data mining is a potent technique that entails extracting valuable patterns, insights, and knowledge from extensive datasets.

1.2 MOTIVATION

Motivation of this study comes from the need for banking institutions to better understand their clients by dividing them into relevant segments. By segmenting customers, banks can tailor their products and services to meet their needs, thereby improving customer satisfaction and loyalty. The use of data mining and predictive modeling techniques may help banks in identifying unique customer segments based on characteristics such as demographics, income, buying behavior, and transactional history.

Customer segmentation can assist Dashen bank in predicting future customer behavior, helping them to make accurate and timely decisions. By identifying patterns in customer spending habits and preferences, banks can anticipate future needs and take proactive steps to address customers. This strategy can effectively result in higher profits by optimizing customer relationships and reducing customer churn.

Ultimately, the motivation of this study is to help Dashen Bank stay ahead of the curve in the ever-changing financial services industry by analyzing customer data, segmenting the clientele, and generating actionable insights, which can be used to improve the bottom line of the banks while increasing customer satisfaction. And also to overcome the manual process of customer segmentation and filtering of customer data, businesses can use automated solutions that streamline and optimize these tasks. By leveraging Data mining techniques, businesses can achieve more accurate and efficient customer segmentation and data filtering processes.

1.3. STATEMENT OF THE PROBLEM

The analysis of customer behavior is currently required for organizations engaged in the banking industry that deal with a large number of clients with diverse characteristics depending on their age, income, financial objectives, and service preferences, such as the need for fundamental checking and savings accounts, interest in loans, interest in limit and guarantee facilities, and interest in digital banking, even though the number of banks and potential clients, as well as the banking services, are increasing. These factors necessitate that banks use data mining techniques to segment their customer base and identify each segment's needs, profile, preferred transactions, and channels. Due to the harsh competition it has produced, banks are still having trouble keeping customers [9]. More people use the banks for daily transactions as the population of the nation rises. People who use banks as a result have more varied needs and anticipated benefits, though these differences do not necessarily indicate that people are fundamentally un-similar.

The following are a few issues that banks face as a result of poor customer segmentation [9]. Accurate information about bank customers is lacking, as is information on how customers behave

during transactions. Due to a lack of market segmentation and customer recognition, advertising costs have increased and its effectiveness has decreased. Additionally, there is a lack of efficient planning, a fragmented approach to acquiring and keeping customers, and an un-segmented class of customers, which prevents the creation of marketing strategies tailored to each segment [9]. Therefore, it is crucial for the banking industry to comprehend consumer behavior and divide consumers into the appropriate groups based on the outcomes. It should be emphasized that customer segmentation enables businesses to gain more insight about customers' behavior in order to satisfy their needs more effectively [2]. At the moment, banks are looking for answers to questions such as which customers are the most preferred customers, who are loyal to the company, and which product may attract more customers. In the decision-making process, customer segmentation is an integral part of the marketing strategy which builds customer relationships, segregates customers into different groups, and provides different facilities in the niche market [6].

In this regard, the problems of Dashen Bank that makes segmentation necessary is that, due to lack of effective customer segregation, the bank does not have a detailed knowledge of its customers need and even customers are not grouped based on their common characteristics. Also Bank do not focus on product offerings and marketing efforts to improve its competitiveness. Besides portfolio management practice is poor that do not optimize customer value through cross-selling and up-selling transactions and pricing. Moreover, the total volume of customer data within the bank makes data comprehension a daunting task.

Belachew [33] conducted a study at BG MFI (Buusaa Gonofa Microfinance Institution) that explored the application of data mining techniques for customer segmentation and prediction. Tenkir [68] conducted a study on the customs management system at ERCA (Ethiopian Revenue and Customs Authority), aiming to evaluate the customs valuation system currently implemented in Ethiopia and identify issues in the practical process, with a focus on the functional process. Belete [34] conducted a study on knowledge discovery for effective customer segmentation, specifically focusing on the case of the Ethiopian Revenues and Customs Authority and identified several gaps, including data integration, limitations in attribute selection, limitations in algorithm selection (particularly clustering algorithms), lower data complexity, and the sum of squared error. All of the mentioned studies primarily employed data mining classification techniques to classify customers.

Hence, this research work is initiated to come up with a data mining techniques that helps to segment and predict customers, so that the Dashen Bank can make proper decisions in designing strategies and looking for additional customers and opportunities to win the market share. This has a significant impact in improving customer relationship management of the Bank.

Therefore, this study attempts to investigate and address the following basic research questions:

- What are the suitable attributes to apply data mining for customer segmentation and prediction?
- What are the suitable data mining algorithms best be used for customer segmentation and prediction?
- To what extent the proposed model works in identifying customer segment?

1.4 OBJECTIVE OF THE STUDY

1.4.1 GENERAL OBJECTIVE

The general objective of the study is to design descriptive and predictive models for Dashen Bank customer segmentation and prediction using data mining algorithms. And to address the manual processing involved in customer segmentation practice.

1.4.2 SPECIFIC OBJECTIVES

To achieve the general objective of the study, the research formulates the following specific objectives: -

- To review related literatures and previous works in the area to have better understanding about the work and identify methods and algorithm for experimentation.
- To collect data on which the mining process is conducted.
- To prepare the data that is used for model building by selecting important attributes, and cleaning them.
- To come up with appropriate number of clusters for customer segmentation based on the similarity of instances.
- To build a predictive model that help in classifying and determining segments of customers.
- To evaluate the performance of the proposed predictive model.

1.4 SCOPE AND LIMITATION OF THE STUDY

The scope of this research is to build data mining descriptive and predictive models. Clustering algorithms are employed for constructing descriptive model towards customer segmentation, i.e. segmenting customers into similar groups. Then using segmentation result, classification algorithms are used for generating hidden knowledge (more specifically rules) for customer prediction.

Dashen Bank start using oracle Flex cube since 2004 and it upgraded the version in 2010 and in 2017. Nowadays the bank is using oracle flex cube 12.2 version for financial and customer data management. Due to the problem of accessing data from legacy system the current research is conducted using 5 years' data which covers 2018-2022 collected from the Banks current CBS Database. The data collected has a content of customer information, saving culture, transaction and credit history.

In this study both descriptive and predictive modeling tasks of data mining are employed. Descriptive modeling using clustering algorithms are employed for grouping the unlabeled data

based on their similarity towards customer segmentation. Predictive modeling is further used to create a model employing classification algorithms towards customer prediction. Accordingly, the study is able to create models and extract hidden knowledge that can help to improve customer relationship management.

In this study, customers who follow conventional banking practices were included due to lack of adequate data on domestic banking practices for Islamic finance (IFB) accounts. Islamic finance operates on principles that differ from conventional banking, and obtaining comprehensive and representative data for IFB accounts may pose challenges.

By focusing on customers who follow conventional banking practices, the study aims to utilize the available data to analyze and draw insights specific to this customer segment. This approach allows the researchers to explore factors, behaviors, and patterns related to the conventional banking sector and draw meaningful conclusions based on the data at hand.

Addition to the above point accounts are selected from 12 branch. And the selection Criteria for Accounts from 12 Branches are

1. **Branch Opening Location:** The location of each branch plays a significant role in determining the customer base and the banking practices followed. By selecting accounts from branches with different opening locations, the study can capture variations in customer behaviors and preferences across different regions or areas.

2. **Branch Opening Purpose:** The purpose behind the opening of each branch can also influence the customer composition and practices. For instance, a branch opened in a commercial area may attract customers with distinct banking needs compared to a branch opened in a residential area. By considering the opening purpose, the study can account for the contextual factors that may impact customer behaviors.

3. **Number of Customers:** The number of customers associated with each branch is an important factor in ensuring a representative sample. By selecting accounts from branches with varying customer counts, the study can encompass a diverse range of customer profiles and behaviors, ensuring a more comprehensive analysis.

The other limitation of this study was the challenge of selecting the most suitable attribute from the vast volume of data. Additionally, out of the initial 13 attributes, only 7 were utilized due to variations in their values.

1.5 SIGNIFICANCE OF THE STUDY

In this study, Customer segmentation and prediction, as a strategic tool helps the bank to work according to the identified segments and the expectations of customers. For bank managers, IT-Data Analytics department and strategy as well as innovation departments this study provides a view to identify customers' preferences, to focus on potential customers who are interested in

Dashen bank product, to improve customer service and help to support decision making. And it also creates an insight for Customer identification which help the banks to increase profitability because services and products that are present in the banks should be based on a better understanding of customers.

Beside this, the study will give an experience for the researcher to conduct research for academic purpose and the finding will give a motivation to conduct further researches in the area. Also the finding of this research used by bank to increase the quality of service given to its customers in order to maintain the standard or the quality of services. This makes customers will be beneficiaries of the quality service provision.

For researchers the finding will give an insight and motivation to conduct further researches in the area. Finally, performing customer segmentation and prediction in banks has advantages for the bank, to gain a competitive advantage and retain customers.

1.6 METHODOLOGY OF THE STUDY

This study uses data mining techniques to segment and forecast Dashen Bank customers. Utilizing research methodology, the research problem can be approached methodically. It might be thought of as a scientific field that focuses on instruction in how to carry out research [6]. This study uses a variety of methodologies to develop the best customer segmentation and classification models for developing and implementing successful customer relationship management. Methods that are descriptive and predictive in particular have been employed. Using descriptive modeling, the behaviors of particular customers have been investigated and clustered into similar groups. Following this, customers are assigned to classes using predictive modeling.

1.6.1 RESEARCH DESIGN

This research follows experimental research. Experimental research methods have a discrete place due to their effectiveness to establish cause-effect relationship and, to make manipulations and provide control over the variables [10].

To conduct an extensive experiment in this research Hybrid Data Mining Process Model is used. According to Swiniarski and Kurgan [11], the hybrid process model is enhanced the knowledge discovery process by combining the academic and industrial models in data mining research. Thus, this model is research-oriented, where its six steps of hybrid models allow a number of feedback mechanisms.

Figure 1.1 below listed the six steps suggested in Hybrid Data Mining process model by Swiniarski and Kurgan [11].

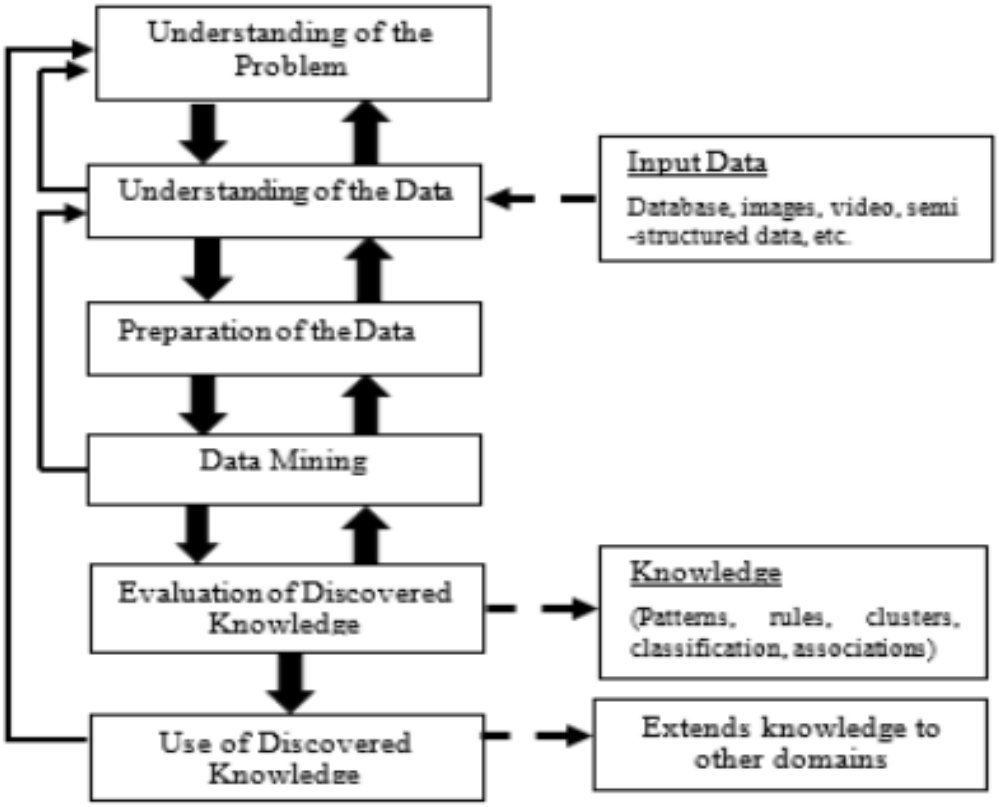

Figure 1.1 Hybrid data mining process model [11]

1.6.2 UNDERSTANDING OF THE PROBLEM

Understanding of the Problem stage works by combining multiple data mining techniques to gain a comprehensive understanding of the problem at hand. This stage involves analyzing the problem domain, identifying the relevant data sources, and determining the objectives and requirements of the data mining process

At this stage of the study tasks such as specifying the problem and goal of the research, identifying important actors, and gaining an understanding of the existing solution by closely interacting with subject-matter experts. Once the study's objectives are translated into data mining (DM) goals, the DM tools to be used in the study are initially selected. The primary aim of the research is to go through several processes in the hybrid model step, which involves examining current customer segmentation, policy and procedure documents, various directives related to customer segmentation from the National Bank, and other domestic banking policies. Therefore, this step is helpful in defining the problem, determining the domain objectives, evaluating the issue and data mining goals, which are useful for the subsequent steps.

Finally, related works are reviewed to have better understanding of the problem, to get the gaps in Dashen bank and how they are working now in world wide.

1.6.3 UNDERSTANDING OF THE DATA

Understanding of the Data stage, a hybrid data mining model works by analyzing and processing the available data to gain insights and understanding. This stage focuses on exploring and preparing the data for further analysis and modeling. During this particular step, the initial data is collected from Dashen Bank's core banking system. The data is then visualized to identify how customers commonly use the bank's products. It is important to ensure that the customer data is complete, as some records may have missing values. The visualization of the data is used to verify whether the data is suitable for the intended data mining goals. Understanding the data goes beyond merely studying and describing it; it also involves identifying potential attributes that can be useful for the purpose of data mining. This stage of the process also involves the selection of appropriate algorithms and techniques for the data mining process. Furthermore, WEKA, which is a tool used for data visualization, can be used to describe the dataset. With its visualization capabilities, the tool can help users gain a clearer understanding of the data and its structure. This understanding is key to ensuring that the data mining process is both efficient and effective, ultimately leading to better insights and decisions.

1.6.4 PREPARATION OF THE DATA

In this step, the data used are prepared to apply the DM methods. It consists of tasks such as sampling, testing the correlation and significance of the data, cleaning the data, checking the completeness of the tuples, handling noisy and missing values. Then, the dimensionality of the data is reduced by feature selection and extraction algorithms. This step also comprises the derivation of new attributes, and summarization of the data. Finally, the datasets that meet the input requirements of DM tools stated in the first step are selected for modeling purpose. It is also including all activities that are needed to construct the final data set. The data which is taken from Dashen Bank core banking database using oracle PL/SQL developer tool is preprocessed and cleaned for the application of different data mining techniques for constructing prediction and descriptive models.

1.6.5 DATA MINING FOR MODELING

This step of hybrid data modeling involves applying data mining techniques to extract valuable insights and patterns from the prepared data. This step focuses on using advanced algorithms and statistical methods to discover hidden relationships, trends, and patterns in the data that can be used for modeling and analysis purposes.In this study, WEKA version 3.8.5 DM software is used. WEKA tool is used because of its ease of use, extensive documentation, and its ability to handle large datasets. It also contains all data mining tasks with possible classification and clustering algorithms for experimentation. Further, it has a range of visualization options, which can help with the interpretation of complex data patterns. It is a knowledge discovery system that provide algorithms for data preprocessing, classification, clustering, and association rules discovery. It also integrates a feature selection and visualization algorithm. For this research work, three clustering algorithms are selected such as K-means clustering, filtered clustering and Farthest First clustering and for descriptive models and tree classification algorithm such as J48 classifier, Naïve Bayes and KNN algorithm for predictive modeling due to convenient platform for this research work.

1.6.6 EVALUATION OF THE DISCOVERED KNOWLEDGE

The evaluation process step involves assessing and validating the insights and patterns discovered through the data mining process. This step focuses on evaluating the quality, relevance, and usefulness of the discovered knowledge to ensure its reliability and applicability in the context of the modeling objectives.

This process involves interpretation of the results by domain experts, checking the impact of the discovered knowledge, and analyzing the evaluation metrics. The primary goal of the evaluation process is to measure and summarize the quality of the trained classifier when tested with unseen data.

One of the most commonly used evaluation metrics is accuracy, which measures the error rate or the generalization ability of the classifier. The accuracy of the trained classifier is measured based on the total number of instances that are correctly predicted when tested with unseen data. The accuracy metric calculates the ratio of correct predictions over the total number of instances that were evaluated.

In addition to accuracy, recall and precision metrics are also used to measure the model's performance and effectiveness. Recall is used to measure the fraction of positive patterns that are correctly classified, while precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class F-score.

Overall, the evaluation process is critical for assessing the quality of the trained classifier and ensuring that it can be effectively applied to real-world problems. By analyzing these evaluation metrics, researchers can gain insights into how well the classifier performs and whether it is suitable for use in different applications.

1.6.7 USE OF THE DISCOVERED KNOWLEDGE

The process of data mining involves the use of various models to extract valuable insights from large amounts of data. Once these models have produced results, the next step is to evaluate their effectiveness. This evaluation process involves assessing whether the new knowledge discovered by the models is novel and interesting.

In this part, several activities are performed to evaluate the results of data mining:

1. **Comparing with existing knowledge:** The data mining team compares the new insights with the existing knowledge in the domain to determine their novelty. If the results are found to be entirely new or significantly different from previous knowledge, they are considered as novel.
2. **Engaging with domain experts:** By collaborates with domain experts to interpret the findings and validate their relevance and usefulness. By cross-checking the insights obtained from the expertise of domain, the researcher ensures the accuracy and applicability of the results.

3. **Assessing interestingness and relationship:** evaluates the interestingness and relationship of the discovered knowledge. This involves determining whether the insights obtained hold significant value and relevance to the business or research problem at hand.
4. **Ensuring validity and usefulness:** The overall evaluation process is crucial to ensure that the insights obtained from data mining are valid and useful for the specific domain. It also ensures that the data mining models used are producing valuable results that can be leveraged to improve decision-making.

Through this comprehensive evaluation process, the researcher aims to identify new knowledge that can drive innovation and enhance decision-making processes within the domain.

1.7 ORGANIZATION OF THE THESIS

This thesis is organized into six chapters that cover a range of topics related to data mining technology and its application in the field of customer relationship management.

Chapter one provides a general overview of the study, including an overview of the background, a statement of the problem, the justification for the study, the objective, the scope, the application, and the methodology of the research. Chapter two is a literature review that covers data mining technology, customer relationship management, and their application. Chapter three delves into the application of data mining techniques and algorithms. Chapter four covers the business understanding of the problem, data understanding, data preparation, data mining, and the evaluation of discovered knowledge, as well as the use of that discovered knowledge. Chapter five focuses on experimentation, with the results of those experiments being analyzed and interpreted in this chapter. Chapter six summarizes the major points of the research and presents the conclusions of the study. Additionally, recommendations have been made for further research in this field.

CHAPTER TWO
LITERATURE REVIEW

The major topics which are indicated in the literature review are understanding of data mining and data mining algorithm. Likewise, previous research work on the problem area was reviewed in order to get the possible applicability of data mining in customer segmentation. Different journal articles, conference papers and the internet publications relating to the subject matter of data mining is reviewed.

2.1 OVERVIEW OF DATA MINING

Data Mining is a multidisciplinary field which supports knowledge workers who try to extract information in the "data rich" environment [12]. The tools it provides assist in the discovery of relevant information through a wide range of data analysis techniques. Any method used to extract hidden patterns from a given data source is considered to be a data mining technique [12].

The process of gaining knowledge from a lot of data is known as data mining. The data could be text, web, time series, spatial, multimedia, and structured transactional data [2]. The process of extracting intriguing, nontrivial, implicit, previously undiscovered, and potentially useful patterns or knowledge from vast amounts of data is known as data mining. To find novel, obscure, unexpected, or unusual patterns in data, a set of activities is used [2]. Data mining is the process of locating pertinent information that enables us to recognize patterns in a vast collection of data. It is very useful for making predictions about the future with the aid of a computer program and is also used in many fields, including biomedicine, gene functions, data analysis of DNA arrangement pattern, disease diagnosis, retail data, telecommunications industry, selling, financial analysis, and astronomy [2]. Data mining algorithms are capable of supporting all of these works.

Due to the widespread availability of enormous amounts of data and the upcoming requirement to transform such data into useful information and knowledge, data mining has recently attracted a lot of attention in the information industry and in society at large. Applications for the information and understanding gained include everything from market analysis, fraud detection, and customer retention to production control and scientific research [12].

2.2 DATA MINING PROCESS MODELS

Process models facilitate the step by step procedure followed during data mining. There are four major Data Mining process models have been used in undertaking data mining projects and researches [18], which include Knowledge discovery in data base (KDD), SEMMA (sample explore modify model assess), CRISP-DM (cross industry standard process for data mining), and hybrid process model.

2.2.1 KNOWLEDG DISCOVERY IN DATABASES PROCESS

Knowledge discovery in databases (KDD) refers to the systematic and iterative process of extracting valuable insights and knowledge from complex and large-scale datasets. This popular data mining technique encompasses several stages, such as data preparation, selection, and cleansing, as well as the integration of previous knowledge pertaining to the datasets. Additionally, KDD involves interpreting accurate and meaningful solutions derived from the observed results, thereby enabling businesses and organizations to make informed decisions based on data-driven insights. [13].

The process of knowledge discovery in databases (KDD) involves multiple steps, with data mining being one essential component. Data mining is the process of extracting new or previously unknown patterns from a large dataset using specific algorithms and techniques. Although the terms KDD and data mining are related, they are not interchangeable. KDD refers to the entire process of discovering knowledge from data, which includes data selection, preprocessing, transformation, and interpretation. Data mining, on the other hand, is a specific aspect of KDD, focusing on the algorithms used to extract useful knowledge from data in databases. Ultimately, the goal of both KDD and data mining is to discover useful patterns or models from data that can be used to inform decision-making in various domains [4]. The knowledge discovery process is iterative and interactive, consisting of six steps (see figure 2.1). Note that the process is iterative at each step, meaning that moving back to previous steps may be required. So it is required to understand the process and the different needs and possibilities in each step [14]. Finally, visualization and knowledge representation techniques are used to present mined knowledge to experts and users in the domain.

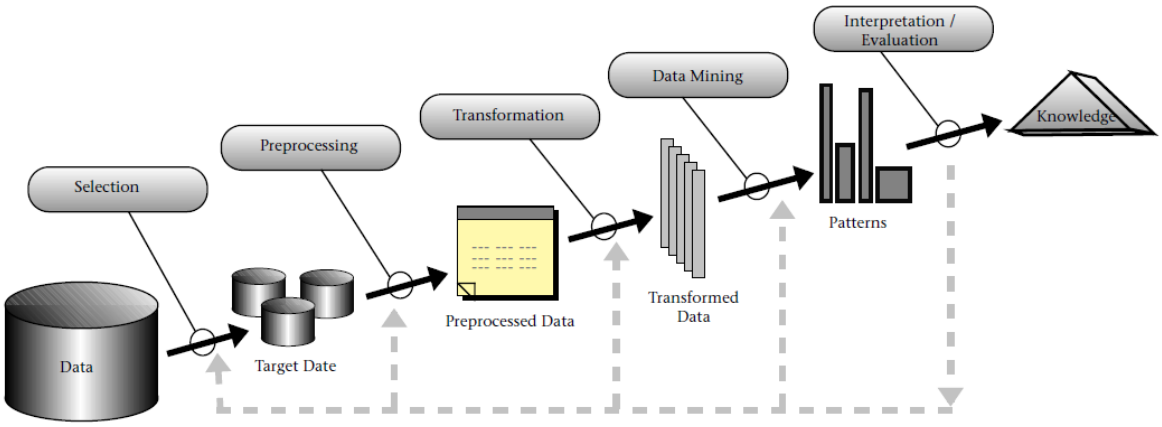

Figure 2.1 Typical knowledge discovery processes [14]

- Data Selection involves analyzing data and identifying the relevant subset of data to be selected for processing. Data sources can be operational data or historical data.
- Data pre-processing involves the process of data cleaning and integration. The data cleaning takes in place filling missing and correcting errors, and integration refers to the

transformation of heterogeneous data extracted from multiple databases into homogeneous data.

- The transformation consists in processing and consolidating data in order to convert them into appropriate formats for the application of data mining algorithms.
- Data Mining is the task of selecting methods to use for searching patterns of interest in particular representational form and apply classification or clustering techniques to obtain predictive and descriptive models.
- Pattern evaluation and Knowledge presentation is where visualization and knowledge representation techniques are used to present mined knowledge to users. It identifies the truly interesting patterns representing knowledge based on Interestingness measures.

2.2.3 THE SEMMA PROCESS MODEL

A data mining process model created by the SAS Institute is called SEMMA, which stands for Sample, Explore, Modify, Model, and Access [66]. It provides and enables the comprehension, organization, creation, and upkeep of data mining projects. It assists in supplying solutions for business problems and objectives. In essence, SEMMA is a logical organization of the functional tools for SAS Enterprise Miner. It has a five-stage cycle, as depicted in figure 2.2 below.

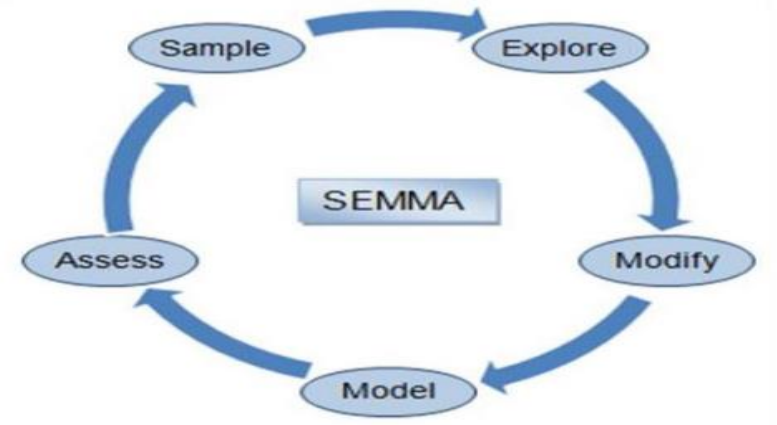

Figure 2.2 SEMMA data mining process model [66]

- **Sample**: is the main focus of this step. A portion of an enormous dataset that is both sizable enough to retrieve significant information and quickly adaptable is extracted.
- **Explore**: Data exploration is the main focus of this step. By searching for patterns and abnormalities, this can aid in understanding and generate ideas while also enhancing the discovery process.
- **Modify**: This step focuses on data manipulation by creating, selecting, and altering variables in order to narrow down the model selection procedure. In addition to looking for outliers, this step might involve lowering the number of variables.

- **Model:** is all about data modeling. This application searches automatically for combinations of data. There are various modeling strategies available, and each type of model is best suited for a particular data mining scenario and has its own advantages.
- **Assess:** This is the last stage of the SEMMA process model, where the reliability and value of the findings are assessed, and performance is estimated.

2.2.4 THE CRISP-DM PROCESS MODEL

Cross-Industry Standard Process for Data Mining (CRISP-DM) provides a uniform framework and guidelines for data miners. It consists of six phases which are well structured and defined [18] as shown below in figure 2.2.

- **Business Understanding**: This is the first phase of CRISP-DM process which focuses on and uncovers important factors including success criteria, business and data mining objectives and requirements as well as business terminologies and technical terms.
- **Data Understanding**: Data understanding concentrates on data collection, quality assurance, and exploration after business understanding is finished in order to gain insight into the data and create hypotheses for hidden information.
- **Data Preparation**: This step focuses on choosing and getting ready for the final data set, taking into account the findings of the data understating. Records, table creation, attribute selection, data cleansing, and data transformation are just a few of the tasks that may fall under this phase.

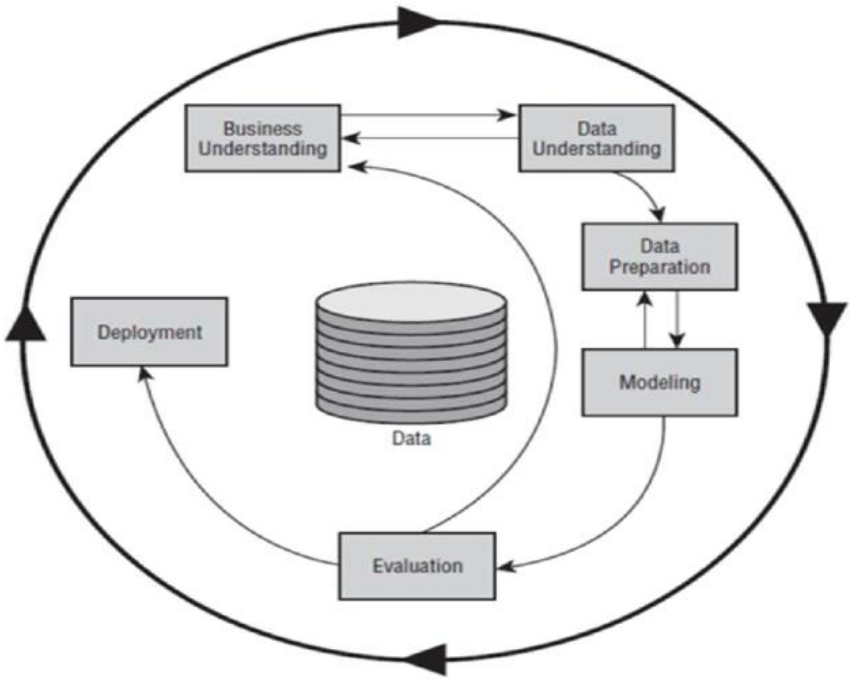

Figure 2.3 CRISP-DM Process Model [18]

- **Modeling:** this is the phase where process selection and application of various modeling algorithm are performed. Different parameters are set and different models are built for same data mining problem**.**
- **Evaluation**: this is the phase which focuses on evaluation of obtained models and deciding of how to use the results. Interpretation of the model depends upon the algorithm and models can be evaluated to review whether achieves the objectives properly or not.
- **Deployment**: This is the final phase of CRISP-DM process which focuses on determining the use of discovered knowledge. This phase also focuses on organizing, reporting and presenting the gained knowledge when needed**.**

2.2.5 HYBRID-DM PROCESS MODEL

Hybrid models, or models that incorporate elements of both KDD and CRISP-DM, have been developed as a result of the development of academic and industrial models. One such model was created based on the CRISP-DM model by applying it to academic research. It is a six-step KDD model. The primary variations and extensions are [11].

- Giving a more comprehensive, research-based explanation of the procedures.
- The modeling step is being replaced with a data mining step.
- A number of new explicit feedback mechanisms are being introduced (the hybrid model has more precise feedback mechanisms, whereas the CRISP-DM model only has three major feedback sources).
- Modification of the final step because, in the hybrid model, knowledge found in one domain may also be used in others.

  The six steps of Hybrid Data Mining Process model, shown in figure 1.1 designed for academic research [11]

2.2.6 COMPARISON OF PROCESS MODELS

As shown by Chapman [59], Table 2.1 present a comparative analysis of KDD, CRISP-DM, SEMMA, and Hybrid DM process Model.

KDD	SEMMA	CRISP-DM	HYBRID
Problem Understanding	Sample	Business Understanding	Understanding of the problem
Selection	Explore	Data Understanding	Understanding of the data
Pre processing	Modify	Data Preparation	Preparation of the data
Transformation			
Data Mining	Model	Modeling	Data mining
Interpretation/Evaluation	Assessment	Evaluation	Evaluation of the discovered knowledge
Post KDD		Deployment	Use of the discovered knowledge

Table 2.1 the difference among DM process model [59]

From the four process model hybrid process model is chosen for this research. The reason behind hybrid Knowledge Discovery Process model lies in its flexibility and adaptability to address the challenges and complexities of data analysis and knowledge extraction. This model allows for iterative cycles, enabling a deeper understanding of the data and the problem at hand with each iteration by following this process, researchers can leverage the knowledge discovered in one domain and apply it to other domains. The six steps of this hybrid model typically include data preparation, data understanding, data preprocessing, modeling, evaluation, and deployment This systematic approach ensures that researchers can make informed decisions and derive valuable insights from the data they are analyzing.

2.3 DATA MINING TASKS

Based on the goals of each task, there are two broad categories into which data mining tasks can be divided. Prediction and description are the main objectives of data mining, according to Fayyad [4] (see figure 2.4). Finding patterns in the data that can be understood by humans is the goal of description, while prediction uses some variables or fields in the database to forecast future or unknown values of other variables of interest. The distinction is helpful for comprehending the overall discovery goal, even though the lines separating description and prediction are not always clear [4] (some predictive models can be descriptive to the extent that they are understandable, and vice versa). For a given data-mining application, prediction and description may or may not be comparatively important. Numerous specific data-mining techniques, including supervised and unsupervised learning algorithms, decision trees, clustering algorithms, neural networks, and association rules, can be used to accomplish the prediction and description goals. For specific data-mining applications, prediction and description can have varying degrees of importance.

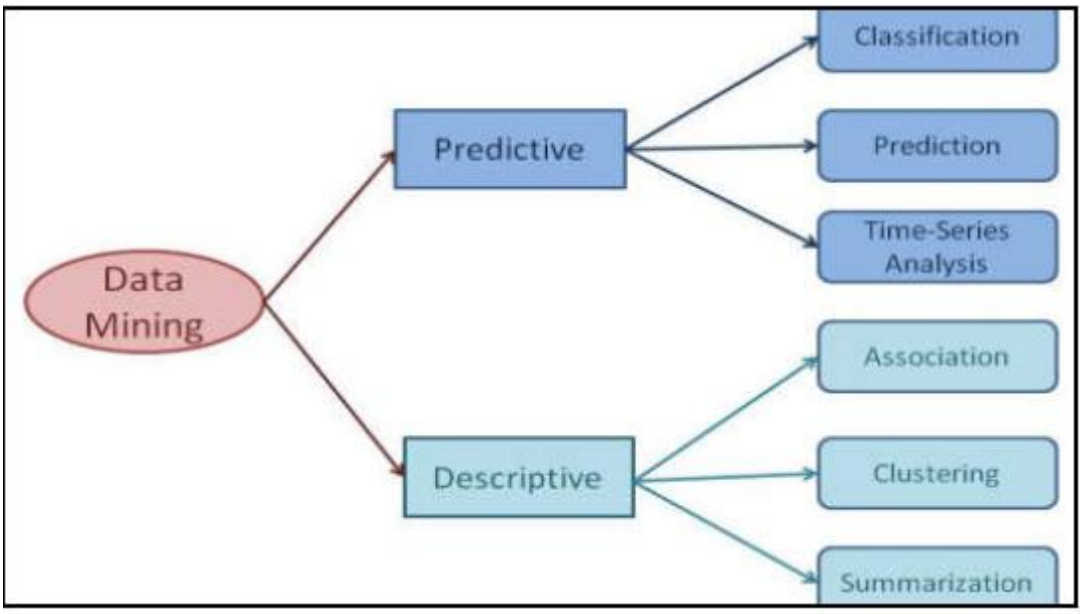

Figure 2.4 Categories of Data Mining tasks [14]

Predictive tasks intended to predict (forecast) a future value or unknown value of a dependent variable (target variable). The model enables to predict the value of one variable from the known values of other variables. It can be used to predict the future habit or behaviors based on the existing or historic data. The predictive task includes classification which is a predictive data mining task aims at predicting (forecasting) a future value or unknown value of a dependent variable and a data type of a dependent variable should be a categorical variable and each value belongs to this variable is called a class label [5]. Building a classification model enables to predict through classifying database records in to a number of predefined classes based on certain criteria. The model is constructed by analyzing the relationship between the attribute and class of the objects in the training set. Such a classification model can be used to classify further objects and develop a better understanding of the classes of the objects in the database.

Descriptive data mining tasks involve analyzing and summarizing the properties of a dataset to gain insights and understand patterns within the data. These tasks focus on describing the data rather than making predictions about future events. Overall, descriptive data mining tasks aim to provide a comprehensive understanding of the dataset by summarizing its properties, identifying patterns, and visualizing the data. These tasks are valuable for exploratory data analysis, data profiling, and generating insights for decision-making.

2.3.1 CLUSTERING ALGORITHMS

Clustering is a valuable method for discovering patterns within data by identifying groups of objects that share similar characteristics. This technique is especially useful when dealing with large datasets where there may not be any clear-cut groups. The main goal of clustering is to group records that are alike. Data mining algorithms that are used to cluster can help to identify natural groupings that may exist within the given data. [37]. Clustering analysis is a technique used to identify groups of data that share similar characteristics. These groups, or clusters, consist of data objects that exhibit common features or behavior. Ideally, a clustering method should produce clusters with high levels of homogeneity within each group and significant differences between groups. This is achieved by minimizing the similarity between clusters and maximizing the similarity within clusters. When conducted effectively, clustering analysis provides insights into patterns and trends in data, which can be useful for a variety of applications. [37]. In other words, in a good clustering result members of a cluster are more like to each other than they are like members of a different cluster. Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build predictive models [37].

Clustering is often confused with classification, but there is some difference between the two. In classification the objects are assigned to pre-defined classes, whereas in clustering the classes are formed. The term “class” is in fact frequently used as synonym to the term “cluster” [38].There are different clustering algorithms, mainly divided in to hierarchical and partitioning algorithms [38].

The task of partition algorithm is to partition a database D containing n objects into k clusters, with the objective of optimizing the chosen partition criterion. Each object is assigned to only one of the k non-overlapping clusters. The k-means algorithm is a type of partition-based clustering method

where the assignment of objects to the k clusters is dependent on the initial centers of these clusters [39].

Hierarchical algorithm creates a hierarchical decomposition of the set of objects either using a top-down approach or bottom-up approach [39]. The agglomerative clustering algorithms use the bottom-up approach and divisive clustering algorithms use a top-down approach. It does not require the number of clusters k as the input but requires a termination condition. Single link and complete link algorithms are examples of the agglomerative hierarchical clustering method. The hierarchical algorithm cannot handle more than a few thousand cases effectively. Thus, sampling the cluster population is required. This task is time-consuming and is not an ideal to sample cluster population. Therefore, it is challenging to apply it for business clustering tasks. Yet, another clustering algorithm called partioning algorithms can handle millions of records without sampling. For this research work, we choose partioning algorithm such as the K-means algorithm, filtered cluster algorithm, and farthest first algorithm.

The centroid-based clustering technique, known as k-means, classifies data into distinct categories without forming a hierarchical structure. The name "k-means" comes from the choice of 'k' clusters and the use of means to calculate a cluster's centroid. This method repeatedly forms clusters by clustering a data set repeatedly. The primary objective of k-means is to precisely identify cluster locations and minimize the distance between the clusters and the data set [43].
The farthest first algorithm belongs to the family of clustering algorithms and is a variation of the K-means algorithm. In this algorithm, each cluster center is placed one after the other at the point that is farthest from the existing cluster center [64]. The first center is selected randomly, and the subsequent centers are chosen greedily as the points that are furthest from the previously selected centers.

On the other hand, filtered clustering technique is used to extract essential information or patterns by filtering out unwanted data. The filtration process is based on specific keywords or relevant information. The algorithm involves storing multidimensional data points in a KD-tree [62], which is a binary tree that uses axis-aligned hyper planes to hierarchically divide the point set's bounding box [63].

2.3.2 CLASSIFICATION ALGORITHMS

Predictive modeling is used to make predictions for new data. A group of predictors that have the potential to influence future outcomes make up predictive models. A model is created once data for these predictors is available. The model could be a straightforward linear equation, a complex neural network, or a decision tree. A model's accuracy is validated or modified using fresh data after it has been created. To estimate or predict the most likely outcome, modeling entails gathering data, creating a predictive model, and applying classification algorithms. In order to analyze and forecast future trends in particular domains, predictive modeling is used. Software usage data, for instance, can be examined to predict future usage trends [17]. Predictive modeling is also applied to live systems in order to evaluate and improve the system's alignment with user and business objectives.

Data mining technique called classification divides the dataset into various classes. The classifiers are the categories [17]. Additionally, it creates a mapping function from a collection of data to help with more precise analysis and forecasting. As the model is used to predict classes of new instances with classification unknown classes, it is used to build models from data with predefined classes. Training data are the instances that were used to build the model [15]. By examining the relationship between the attributes and the classes of the objects in the training set, a classification function or model is built. It is possible to classify new objects and gain a better understanding of the classes of the objects in the database by using a classification function or model like this one. The major goal of a classification algorithm is to maximize the predictive accuracy obtained by the classification model. There are different Classification Algorithms available, here under the well-known and common algorithms are discussed.

In data mining and knowledge discovery, decision trees are the most effective method. It involves using technology to analyze vast, complex amounts of data in order to find patterns that are helpful. The ability to model and extract knowledge from the vast amount of available data makes this concept extremely significant. Every expert and theorist is always looking for ways to improve the procedure's accuracy, cost-effectiveness, and efficiency [47]. In many fields, including information extraction, machine learning, pattern recognition, data and text mining, decision trees are very useful tools.

The Naive Bayes classification algorithm is very straightforward and makes the assumption that the classification attributes are independent and do not interact in any way. In many cases where other alternative methods are suggested to improve performance, researchers have discovered that this assumption of independence is false. Based on conditional probability and maximum likelihood occurrence, the original Naive Bayesian technique [51] is used.

Among the supervised learning algorithms, the K-nearest neighbor classifier is a crucial one. A new set of data is categorized using this method based on how many votes its nearest neighbor has received. A distance function that uses the Euclidean, Manhattan, and Minkowski distance methods to calculate the distance between nearest neighbors. When determining the class of a new instance, the K value denotes the number of neighbors that were taken into account. One will find results that are less stable if the K value is very low. However, one will still get stable results by increasing the K value [54]. On the other hand, doing so will allow for an increase in error.

2.4 DATA MINING APPLICATIONS

According to Kpal [60], discussion data mining has a great contribution in solving business problems by finding patterns, associations and correlations which are hidden in the business information stored in the data bases. Some of the common application areas of data mining are as follows.

- **Customer Relationship Management**: Data mining technique can be used to create customer profiling to group the likeminded customers in to one group and hence they can be dealt accordingly [19].
- **Fraud Detection**: Customers and banks both run the risk of becoming easy targets for frauds when dealing with banks. Therefore, both parties want to feel safe when dealing with one another. They can identify frauds and subsequently prevent them with the aid of data mining techniques. The organization will be able to concentrate on ways and means of analyzing customer data in order to identify patterns that may lead to frauds with the aid of data mining techniques [20].
- **Risk management:** Data mining techniques are used to quantify the risk involved in credit facility decisions, simplifying the process and limiting any financial losses to the bank at the same time [21].
- **Prediction**: As its name suggests, the prediction is one of the data mining techniques used to identify relationships between independent variables and those between dependent and independent variables [2].
- **Money Laundering Detection:** Money laundering is the practice of disguising the illegal source of "black money" in order to give it legitimacy [36]. Money laundering frequently takes place through banks. Governments and financial regulators therefore mandate that banks implement processes, systems, and procedures to identify and stop money laundering activities.

2.5 SIGNIFICANCE OF CUSTOMER SEGMENTATION AND PREDICTION

Customers can have different types of characteristics and can be of different importance to a company. For companies to know the characteristics of customer, segmentation of customers need to be done [22]. Customer segmentation allows us to better match customers with products that are similar to their preferences. By analyzing customer data, this approach changes the way we communicate with customers to ensure we are providing them with relevant offers. Additionally, it helps us identify the most profitable customers and enables us to update our products and services to better meet their needs. [23].

Segmentation theory refers to the practice of categorizing customers into different groups based on their unique characteristics. One common approach that companies use for segmentation is to analyze how much revenue each customer contributes to the company based on their purchase volume. This enables businesses to better understand their customers and tailor their marketing strategies and product offerings to meet the specific needs and preferences of each group [24].

By identifying and categorizing customers, we can gain a deeper understanding of their needs and the type of demand they require. Certain customer segments may have a high level of innovation, leading to frequent changes within the group over time. To effectively serve these customers, it is vital to stay aware of these shifting requirements and adapt to meet their needs in the most effective and efficient way possible. This approach will help ensure we are meeting the evolving needs of our customers and providing the highest level of service possible. [25].

When a company is well-informed about the requirements of its clients, it will be easy for it to divide those clients into various groups. Additionally, it will be simpler for the company to identify what surprises and even delights its clients. Such information can be used to improve their services or goods in the future. It is crucial for businesses to have this component in place because today's consumers value customer service on par with the actual product or service. By segmenting customers, it is easier to decide how much and what the company should emphasize when determining the level of services that the different groups should receive [26]. Retaining highly profitable customers in today's fiercely competitive business environment is the main challenge, but it is possible to do so by steadily improving customer-centric products and services. Understanding a customer is the first step in building a strong relationship with them [15]. According to Lambert [27], market segmentation is essential in both the production and service sectors that are developing. All kinds of businesses need to figure out the best strategy for breaking the market down into distinct groups in order to effectively meet customer demand and increase revenue. An organization provides a product or service to its clients. In order to satisfy its customers, a company must therefore adhere to the customer service standards [28]. What a company does to involve customers, vendors, and other parties.

2.6 RELATED WORKS

The literature review is done to get more information about the problem and to know which type of works done related to the current study and to know the methods and algorithm other research used with the result they achieved and the way forward recommended.

Koyuncugil and Ozgulbas [31], from Turkey, have explored the possibility of designing data mining techniques for financial institutions using data obtained by means of financial analyses of balance sheets and income statements of companies under Turkish Central Bank. They came up with a model for detecting financial and operational risk indicators of Small and Medium Enterprises (SMEs). The research aimed to develop a customer profiling segmentation model using the decision tree algorithm, specifically employing the CHAID (Chi-Square Automatic Interaction) method in their study. According to the findings, this approach resulted in the creation of an accurate model capable of identifying financial and operational risk indicators in SMEs. The researchers suggest that future data mining studies could make use of customer data and explore additional classification techniques (such as decision trees and neural networks) to achieve various mining objectives.

An analysis was conducted by Begunca, [32] to study the behavior of consumers in market segmentation based on benefits sought or required. The primary aim of this study was to identify behavioral features and characteristics based on the benefit sought approach, to assist in the development of effective segmentation strategies. The researcher employed both quantitative and qualitative methods for this study. The quantitative approach facilitated statistical data analysis to uncover relevant information, and the k-means cluster analysis helped identify relatively homogeneous groups of cases based on the selected characteristics. The customer base was divided into five clusters, each with its own unique attributes.

- Cluster 1: This segment of customers prioritizes consumption factors, followed by sports and entertainment factors like energy and entertainment. Additionally, the freshness of the drinks is also important to this group.
- Cluster 2: This segment prioritizes sports and entertainment factors first, followed by freshness and consumption.
- Cluster 3: This segment prioritizes health aspects like low calories, natural ingredients, and vitamins, followed by freshness and symbol status factors like image, quality, and famous brands.
- Cluster 4: The fourth segment prioritizes freshness aspects like carbonation, freshness, and caffeine, followed by consumption factor and symbol status factor.
- Cluster 5: The final segment gives more importance to the symbol status of the drink, including factors like image, quality, and famous brands, followed by consumption and freshness factors.

Furthermore, the researcher concluded that there is a need to segment consumers based on the specific benefits they require or seek. This information can help in creating effective marketing strategies that cater to the needs and preferences of each segment.

Aghaei [9] delved into data mining techniques to better understand customer behavior at Shahr Bank of Iran, which has faced several challenges in terms of poor customer service and orientation. Amongst the problems faced by the bank are the non-segregation of customer data, lack of an effective marketing plan, and high marketing costs. These challenges deemed it necessary to identify and segment customers. Using factor analysis, cluster analysis, and conceptual maps analyzed using SPSS 24, the study aimed to segment Shahr Bank customers into four categories, including benefit-oriented, peace-oriented, interest-oriented, and moderate. The study revealed that the two categories with the highest percentage of male customers were peace-oriented (69.0%) and moderate (65.6%), while singles made up the majority of peace-oriented customers (51.6%) and qualified for the majority of moderate customers (52.2%). Additionally, both peace-oriented (38.1%) and moderate (39.4%) categories had the highest percentage of customers within the 25 to 35 years' age group. Finally, the peace-oriented customers had a relationship period ranging from 5 to 10 years (31.7%), while the moderate customers had a relationship period of 1 to 5 years (41.3%).

Shahenaj [61] His study looked at customer transaction patterns, product holdings, demographics, and historical trends to segment and profile customers. K mean clustering algorithm and scorecard were used in the study to segment and profile customers. Continuous valued variables are used by the researcher for analysis. If a customer falls into the first quartile, they receive a score of 1, for the second quartile they receive a score of 2, for the third quartile they receive a score of 3, and for the fourth they receive a score of 4, respectively. If a customer has a credit card, they would receive a score of 1, otherwise they would receive a score of 0, and each variable would be given a weight between -3 and +3. The researcher divided them into 4 quartiles according to the total score bracket. The highest scoring category was Q4, and the lowest scoring category was Q1. Q2 and Q3 are still in the middle. A customer would be profitable if their billing period fell in the fourth quarter. This would assist the bank in dividing up its customer base into different groups according to the profitability brought about by client relationships. The most profitable, least profitable, and

profitable customers were divided into three clusters by the researcher after analyzing the final score. Out of a base of 4500 customers, the researcher identified 3184 as the potential and profitable customers who add value to the bank's profit (falling in Q2, Q3, and Q4). the prohibition.

In addition to foreign works, there are local works done to apply data mining in customer segmentation and profitable customer prediction.

Belachew [33] explored the likelihood of applying data mining techniques for predicting profitable customers and segmenting the existing customer to identify best customer segments from bad customer segments. The researcher focused on building a model that helps to classify customers for product and services offering by Buusaa Gonofa Microfinance Institution using a predictive data mining model. To classify customers in the organization there is no data with predefined classes. The researcher; therefore apply clustering techniques that result in an appropriate number of clusters. The study was done on the data set of BG MFI (Buusaa Gonofa Microfinance Institution) customer data. In this work, Weka version 3.7.5 is selected to implement data mining. The methodology of this research consists of four stages such as data collection, data preprocessing, data mining (classification and clustering) and interpretation. K-means clustering techniques was used to cluster the data set into five groups. Based on the result obtained through clustering, decision Tree (J48) algorithm is employed for classification using a J48 decision tree algorithm the researcher conducted different experiments out of which model constructed by 10-fold cross-validation test model performs better with accuracy of 99.95%. In general, this study helps to identify the potential customer for an institution for better customer relationship management. Finally, the researcher recommended that experimental tests be conducted by the institution with inclusion of many datasets by using large training and testing datasets, development of an integrated data warehouse, development customer relationship management strategies.

Belete [34] explored how to design a model using data mining techniques for customer segmentation that helps Ethiopian Revenue and Customs Authority treat customers according to their behavior and how to increase the revenue of the organization. The researcher uses quantitative methods to collect and analyze customer's data and qualitative methods to understand the business operation. The researcher used KDD (Knowledge Discovery in Database) process models to produce useful patterns or models. In this study, clustering K-means techniques were used to cluster the subset of the dataset into five groups Very High, High, Average, Low, and Very Low. Further classification techniques are applied to predict future customer's behaviors. For classification J48 decision tree algorithm and multilayer-perceptron algorithms had been experimented. After conducting various experiments, the researcher conclude that 10-fold cross-validation has better classification with 99.95% accuracy than percentage split experimentation. Finally, the researcher concludes that the J48 tree algorithm is the best algorithm to classify the customer value and the clustering algorithm is used for customer segmentation and also used to identify the characteristics of the customers. The study recommended that Customer Data Warehouse need to be build, Inclusion of additional customer attributes are needed for better customer segmentation.

Fikrealem [35] developed clustering models that identify the behavior of high-value enterprise customer's using data mining techniques. The data mining tasks conducted based on the six steps of hybrid data mining process model, such as, understanding of the problem, understanding of the data, preparation of the data, data mining, evaluation of discovered knowledge and use of

discovered knowledge. The study conducted using WEKA software version 3.8.2 and three data mining algorithms, such as k-means, filtered cluster and farthest first. After conducting various experiment, the researcher concludes farthest first algorithm with cluster size (k) 2 and seed size 100 has a better clustering performance. It enables to cluster corporate customers into a dissimilar cluster of high and low-value customer's groups of the corporation and also enables to identify customer behavior in each cluster group. The result of the research shows that applying data mining helps Ethio-Telecom to know enterprise customer value and behaviors, to understand consumption characteristics of different customer groups, to identify target customer groups, characterizes each group and analyze their properties. As a recommendation the researcher forwards for further study by adding residential customer data, regional enterprise customer data, and call duration, times of call, the number of different telephone numbers called by the speaker, concentration of call duration and concentration of times of calls.

Related research review shows that there are few research works in customer segmentation and prediction on local banks by applying data mining and machine learning algorithms. Hence this study aims to apply data mining for customer segmentation and prediction and also to come up with hidden patterns and knowledge that can help Dashen bank to improve its CRM.

To summarize both local and foreign works Koyuncugil and Ozgulbas' research in Turkey developed a customer profiling segmentation model using the decision tree algorithm to detect financial and operational risk indicators in SMEs. Begunca's study analyzed consumer behavior and identified five distinct customer clusters based on consumption factors, sports, entertainment, health aspects, freshness, and symbol status. Aghaei's study at Shahr Bank of Iran segmented customers into four categories based on orientation, age, and relationship periods. Shahenaj's study used K-means clustering and a scorecard approach to segment customers into profitability groups. Belachew's study predicted profitable customers and segmented existing customers for Buusaa Gonofa Microfinance Institution, achieving a high accuracy rate of 99.95%. Belte's study at the Ethiopian Revenue and Customs Authority used clustering techniques and classification algorithms to predict customer behavior and segment customers.

CHAPTER THREE

METHDOLOGY OF THE STUDY

3.1 OVERVIEW

Data mining encompasses a variety of techniques, such as clustering, classification, association rules, and regression analysis. Among these techniques, clustering and classification are often used in customer segmentation and prediction. Clustering algorithms seek to identify meaningful groups of data based on similarities between data points. These algorithms also calculate the centroid or center point for each cluster. To assign data points to specific clusters, most algorithms measure the distance between a point and the centroid of each cluster. The output of a clustering algorithm is typically a statistical summary of the centroids and the number of data points in each cluster [41].

Data mining frequently utilizes classification, which is a crucial technique that involves identifying a mapping function to classify data into pre-defined categories or concepts. However, in this particular study, the data collected didn't have any pre-existing categories that described the institution's clients. Thus, the researcher instead employed clustering techniques to generate the optimal number of clusters. The subsequent section delves into the data mining methods and algorithms utilized in this research project.

3.2. THE PROPOSED ARCHITECTURE

The major components of any data mining system architecture are the guideline of the data modeling for prediction, i.e. data source, data mining engine and pattern evaluation module. The data mining process identify effective model for customer segmentation and prediction. It is divided into six steps. The processing blocks in the proposed architecture are shown in Figure 3.1.

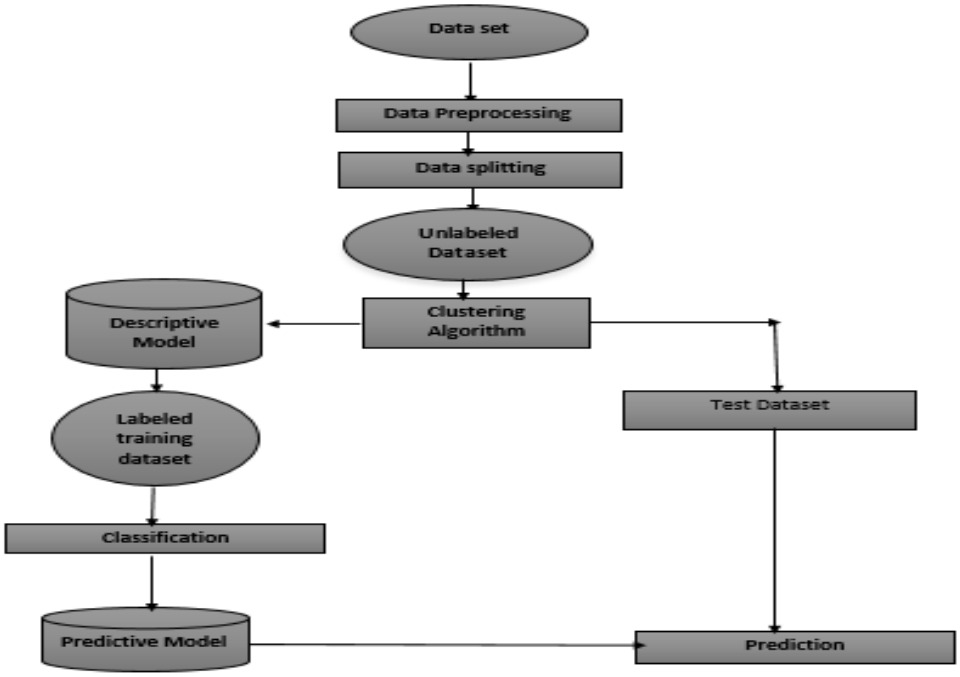

Figure 3.1 The proposed architecture for Customer segmentation and prediction.

As shown in the architecture,

Dataset: The initial dataset was collected from the banks' CBS, with a total size of 45,698 records. The collected Data has demographic, psychographic financial variables and behavioral data.

Data preparation: The data preprocessing stage is crucial for data analysis. In this stage data cleaning, data integration, transformation and smoothing have been performed. In Data cleaning missed values are filled with standard values. The Data set is collected by integrating from different CBS tables. In transformation operation column name of FIELD_VAL_17 and total debit and credit columns are changed to Initial Deposit, Monthly_credit_tov respectively

Unlabeled data: In the unlabeled data, a clustering algorithm K-Mean, Farthest first and Filtered was applied to obtain a descriptive model that assigned labels to the dataset.

Data split: The dataset was divided into training and test sets, with 70% (31,988 records) allocated for training and 13,710 records for testing purposes.

This labeled dataset was then used to apply a classification algorithm which is Decision tree, KNN and Naïve Bayes resulting in a predictive model.

3.3 CLUSTERING ALGORITHMS

The process of clustering is a common technique in unsupervised learning, which involves grouping data points based on their similarities. Using a clustering algorithm, large numbers of data points are assigned to smaller groups, ensuring that like-minded data points are grouped together while dissimilar points are kept apart. This technique is crucial in data mining as it enables the identification of patterns that may be difficult to detect through other methods. Clustering is especially significant in multivariate applications such as market forecasting and planning research, as it allows for better segmentation and analysis of data. [42].

Clustering is a crucial aspect of numerous data mining applications, including scientific data exploration, information retrieval, text mining, spatial database applications, web analysis, customer relationship management (CRM), marketing, medical diagnostics, computational biology and various other fields. [42]. In this study k-means clustering, filtered clustering and Farthest First clustering algorithms that are experimented for bank customer segmentation.

3.3.1 K-MEANS CLUSTERING TECHNIQUE

One of the most popular unsupervised clustering techniques is K-Means Clustering [43], which utilizes a centroid-based approach to divide data into non-hierarchical categories. The name "k-means" arises from the fact that the letter k represents the number of clusters chosen, and "mean" refers to the method used to calculate the cluster centroid. This iterative algorithm can be applied repeatedly to a dataset, forming clusters and determining their locations by minimizing the distance between the cluster and the data points. Ultimately, the primary objective of this algorithm is to effectively cluster the dataset [43].

In K-Means, the centroids are computed as the arithmetic mean of similarity/dissimilarity all points of a cluster. The distances are computed according to a given distance measure, e.g. Euclidean distance.

K-means has two significant disadvantages despite having the great benefit of being simple to use. First of all, because each step requires calculating the distance between each point and each cluster, which can be very costly when there is a large dataset, it can be very slow. Second, this method is extremely sensitive to the initial clusters provided; however, this issue has been somewhat resolved in recent years [44].

Here under presented the step by step procedure followed by k-means clustering algorithm [44].

1. Define the number of clusters (k) to be produced and identical data point centroids.
2. The distance from every data point to all the centroids are calculated and the point is assigned to the cluster with a minimum distance.
3. Form the set K clusters by assigning each data point to the centroid that is closest to it.
4. Calculate the variance and move the centroid of each cluster.
5. Reverse the previous three steps, reassigning each data point to the cluster's new closest centroid.
6. Go to step 4 if there is a reassignment; otherwise, go to step 7.

7. The algorithm is now stop.

Figure 3.2 below shows summary of the flow of k-means clustering algorithm.

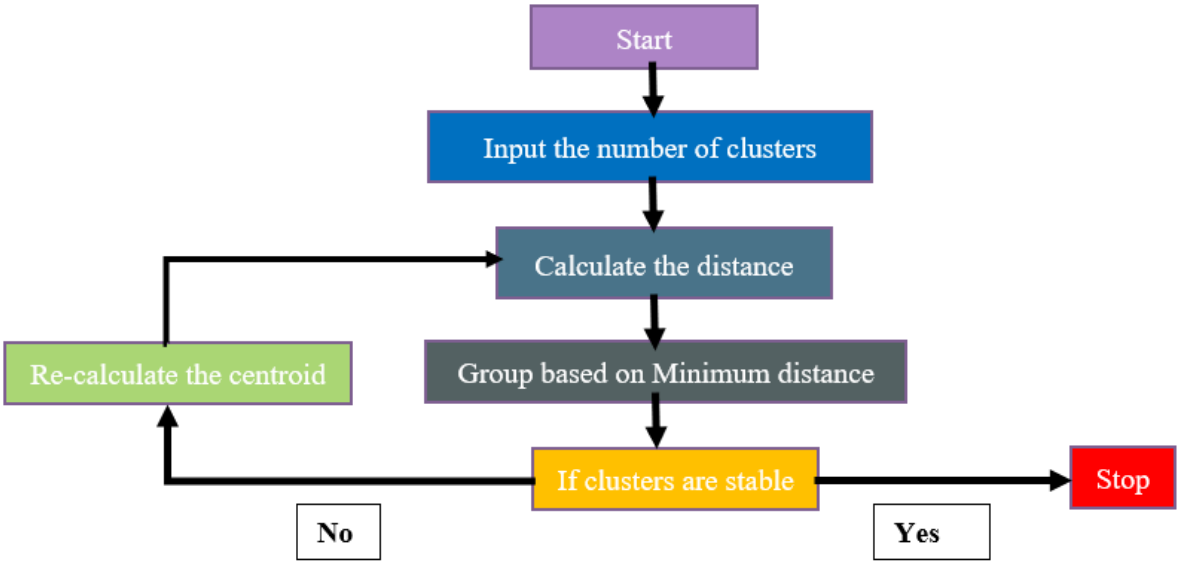

Fig 3.2 Flow of K-means algorithm [42]

After the clusters have been created the result should be interpreted. According to Berry and Linoff [45], the three commonly used approaches to understand clusters or class are:

1. Examining the differences in the distributions of variables from cluster to cluster, one variable at a time.
2. Using visualization to see how the clusters are influenced by changes in the input variables.
3. Building a decision tree with the customer label as the target variable and using it to generate rules explaining how to assign new records to the correct cluster. Using the above three approaches interpretation of the output was performed to have better understanding on customers clustering.

3.3.2 FILTERED CLUSTERING TECHNIQUE

The Filtered Clustering algorithm is designed to extract specific patterns or information from a dataset by utilizing provided keywords or relevant data points. This filtration process is accomplished through the use of a kd-tree, in which multidimensional data points are stored. A kd-tree is a binary tree structure that hierarchically subdivides a set of points within a bounding box by using hyper planes that are aligned with the axes of the dimensions of the points [52][63]. By utilizing this tree-based approach, the Filtered Clustering algorithm can efficiently and effectively filter through large datasets to extract only the information that is necessary for analysis or further processing. The following algorithm shows how filtered cluster algorithm [63].

1 First, compute the candidate that is closest to the midpoint of C.

2 Then, for each of the remaining candidates, if no part of C is closer to z than it is to z*, it infer that z is not the nearest center to any data point associated with u and

3 Hence, it prune, or "filter" z from the list of candidates. If u is associated with a single candidate (which must be z*) then z* is the nearest neighbor of all its data points. It assign them to z* by adding the associated weighted centroid and counts to z*.

Summary of the Filtered Clustering step by step procedure is presented in figure 3.2 below.

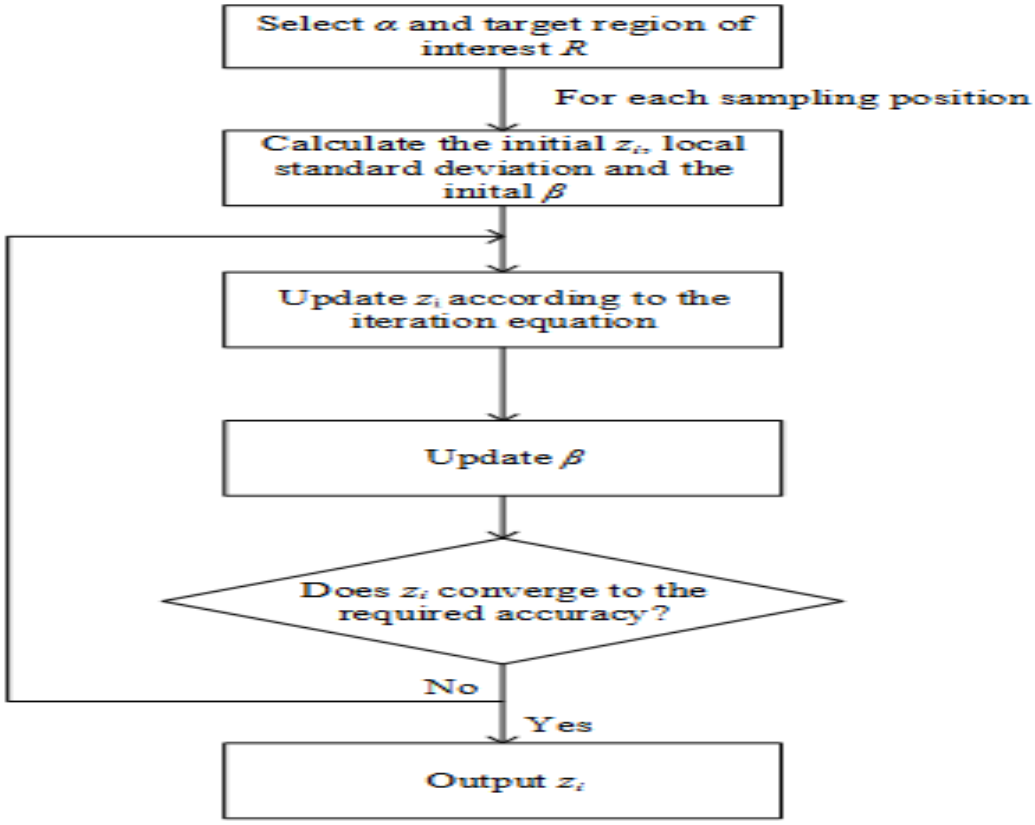

Figure .3.3 the Filtered Clustering Algorithm [52]

3.3.3. FARTHEST FIRST ALGORITHM

The farthest first algorithm is one of clustering algorithm; it is a modification of K-means that places the center of each cluster in turn at the point furthermost from the existing cluster center [64]. The first center selected randomly and the second center is greedily selected as the points further from the first. The following step used to cluster a given data set D using the farthest first clustering algorithm [65]:

Step 1: Randomly select first center
Step 2: For (i= 2…, k) do // select centers
Step 3: For (each remaining point) do
Step 4: calculate distance to the current center set;
Step 5: Select the point with maximum distance as new center
Step 6: Apply Euclidean Distance function on each cluster
Step 7: for (each remaining point) do //assign remaining points
Step 8: Calculate the distance to each cluster center using Manhattan distance formula.

Step 9: put it to the cluster with minimum distance
Step 10: repeat the steps until each cluster remain the same

Hereunder figure 3.4 presents the result returned by farthest first clustering algorithm.

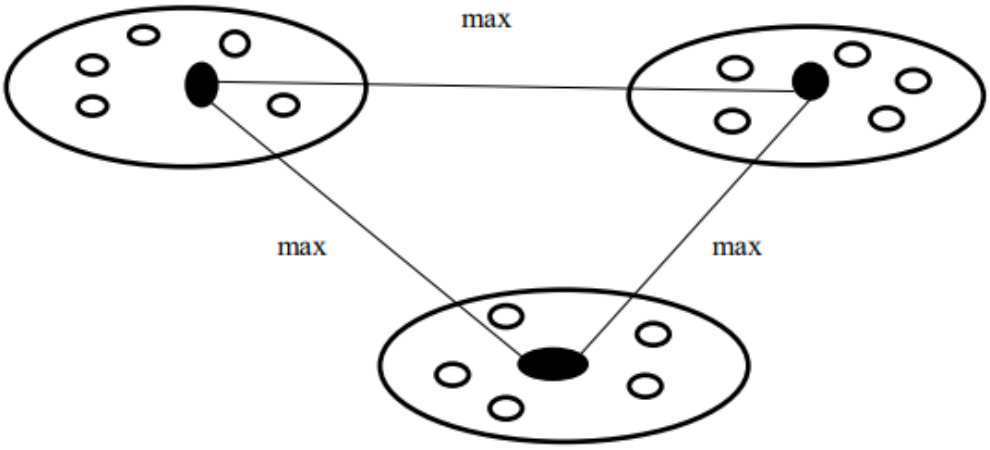

Figure 3.4 Farthest first clustering Algorithm [65]

3.4 CLASSIFICATION ALGORITHMS

Classification refers to the process of identifying patterns and developing a set of models or predefined conditions that can describe and differentiate various data classes or concepts. This is a supervised learning method, which means that it relies on a collection of labeled patterns to learn the descriptions of the different classes. Once these descriptions have been learned, they can then be used to label or classify newly encountered, yet unlabeled patterns. In other words, the goal of classification is to use the knowledge gained from the available labeled patterns to accurately label new patterns so that they can be appropriately categorized.

Classification is a method of grouping data into predetermined categories. Different types of models, such as decision trees, neural networks, Bayesian networks, and "If-then" rules, can be used to represent the derived classification model. Usually, this technique uses decision tree or neural network-based classification algorithms. The classification process involves two stages: "learning" and "classification." In the learning stage, the classification algorithm analyzes the training data. In the classification stage, test data is utilized to evaluate the accuracy of the classification rules. If the accuracy is deemed acceptable, the rules can be applied to new data tuples [46]. In this study decision tree, k-nearest and naïve Bayes classification algorithms are experimented for bank customer prediction based on the result of descriptive model constructed employing clustering algorithms.

3.4.1. DECISION TREE

The use of decision trees is one of the most powerful approaches in the field of knowledge discovery and data mining. By sifting through large volumes of data, decision trees can uncover

useful patterns that would otherwise remain hidden. This is a crucial concept, as it allows for modeling and extracting knowledge from vast amounts of information. Professionals in the field are consistently searching for ways to improve the efficiency, cost-effectiveness, and accuracy of this process [47]. Decision trees are highly effective tools in various areas, including data and text mining, information extraction, machine learning, and pattern recognition. Decision tree offers many benefits to data mining, some of them are as follows [47].

- It is easy to understand by the end user.
- It can handle a variety of input data: Nominal, Numeric and Textual.
- It is able to process erroneous datasets or missing values.
- It registers high performance with small number of efforts

A decision tree's structure is similar to a tree, with leaf nodes at the end designating the classes to which the data is assigned, internal nodes standing in for tests on attributes, branches for test outcomes, and internal nodes for tests themselves. In a tree, the root node is the node that is highest up. Every node could have two or more branches, depending on the algorithm. For instance, the CART (Classification and Regression Trees) algorithm creates trees with just two branches at each node. The term "binary tree" refers to such a tree. The term "multi-way tree" [47] refers to a tree that has more than two branches at each node. Decision trees are constructed using top-down recursive, divide-and-conquer methods. To represent all training datasets, the tree's root node is at the top. [48].

1. If the training lists have the same outcome, the node will be leaf and it is labelled with that class.
2. Otherwise, the tree selects the greatest information attribute to divide the set and labelled the node by the name of the attribute.
3. Recur the steps and stop when all samples have the same class or there is no more samples or new attributes to portion.

Summary of the steps followed in decision tree construction for customer prediction is depicted in figure 3.5.

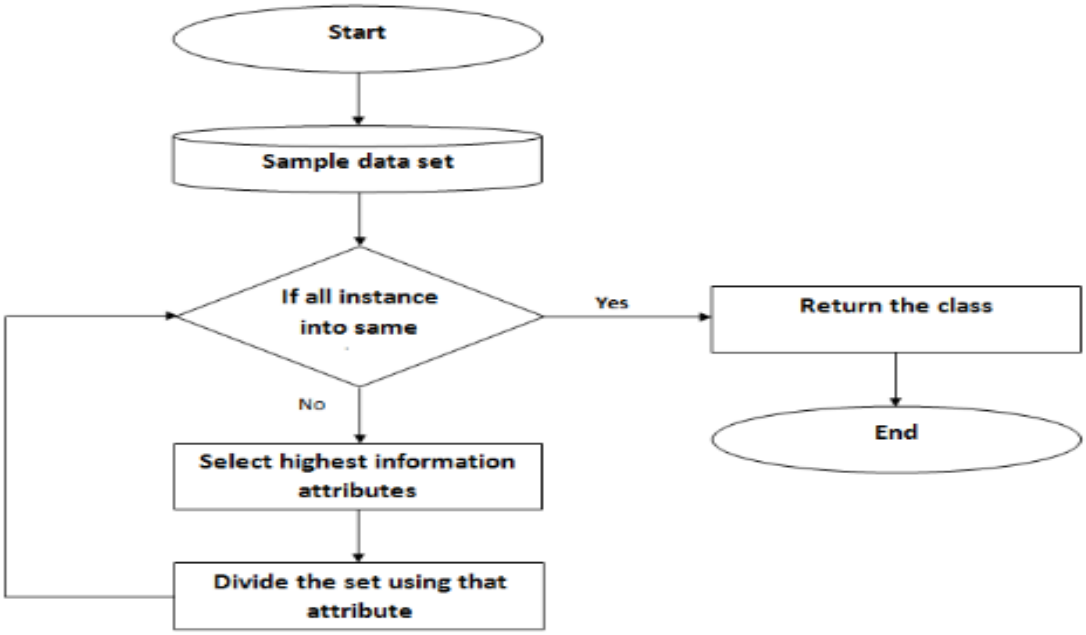

Figure 3.5 steps in Decision Tree construction [48]

3.4.2 NAÏVE BAYES

Naive Bayes is a data classification algorithm that is considered to be one of the top 10 algorithms in data mining [49]. The Naive Bayes algorithm is a straightforward approach to probabilistic classification. It works by computing a set of probabilities based on the frequency and combination of values present in a given data set [50]. To obtain the probability of specific features in the data, they are listed as part of a probability sequence, which is derived from the frequency of each feature value in the corresponding class of the training data set. This set, which is used to train classification algorithms, contains a subset of the available data and is employed to predict unknown values based on known examples [51].

The very basic Naive Bayesian classification algorithm makes the assumption that each classification attribute is independent of the others and that there is no correlation between them. Numerous researchers have discovered that this independence assumption is not true in all situations where other alternative techniques are suggested to improve performance. Based on conditional probability and maximum likelihood occurrence, the original Naive Bayesian method is used. Here is the step-by-step procedure followed by Naïve Bayesian Algorithm [51].

1. For each class calculate the probability of the given instance not belonging to it.
2. After calculation for all the classes, check all the calculated values and select the smallest value.
3. The smallest value (lowest probability) is selected because it is the lowest probability that it is NOT belong to that particular class. This implies that it has the highest probability to actually belong to that class. So this class is selected.

Figure 3.6 below provides summary of the steps followed in Naïve Bayes classification algorithm for customer prediction.

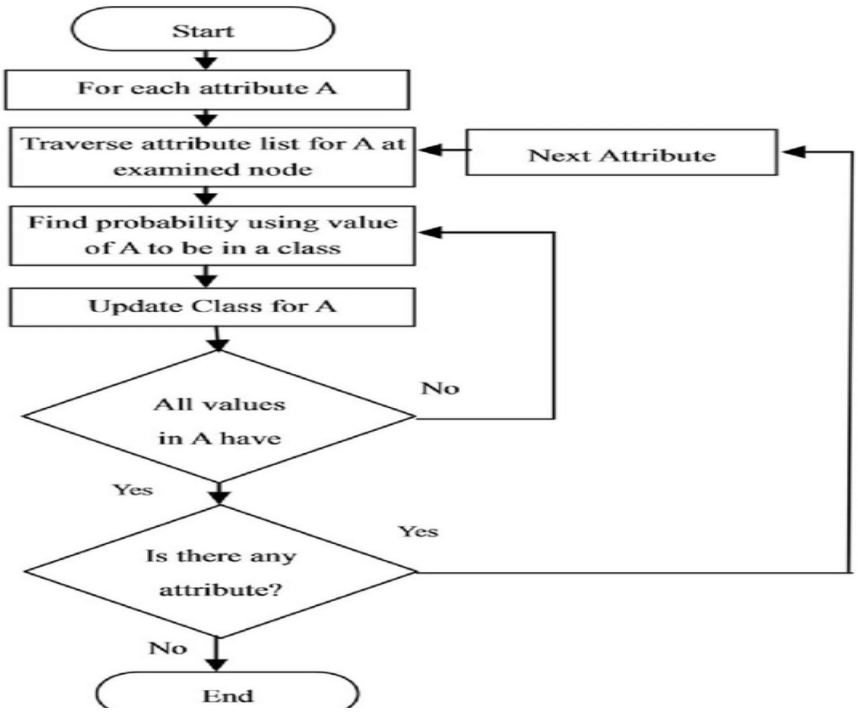

Figure 3.6 Naive Bayesian algorithm [55]

3.4.3 K-NEAREST NEIGHBOR

K-nearest neighbor is a supervised learning algorithm that is widely used for classification. The main idea behind this method is to classify a new data point based on the class labels of its K-nearest neighbors. The distance between the new data point and its neighbors is calculated using a distance function. The most commonly used distance functions are Euclidean distance, Manhattan distance, and Minkowski distance.

The value of K determines the number of neighbors that will be used to classify the new data point. If K is too small, the model will be sensitive to noise and outliers, leading to less stable results. On the other hand, if K is too large, the model will become less sensitive to local variations and will tend to classify new data points based on the most frequent class label in the dataset. This is known as the bias-variance trade-off.

In summary, K-nearest neighbor is an important classifier in supervised learning that relies on the distance between the new data point and its neighbors to determine its class label. The choice of K is critical as it affects the stability and accuracy of the results. By selecting an appropriate K value, one can balance the bias-variance trade-off and obtain stable and accurate results.
Here under step-by-step procedure is present for KNN Algorithm [54].

1. Determine the parameter K
2. Calculate the distance between the data to be evaluated with all the training data
3. Sort range formed (in ascending order)
4. Determine the shortest distance to the order of K

5. Pair the corresponding class
6. Find the number of classes from the nearest neighbor and set the class as a class data to be evaluated

Figure 3.7 below provides summary of the steps followed in KNN classification algorithm for Customer prediction.

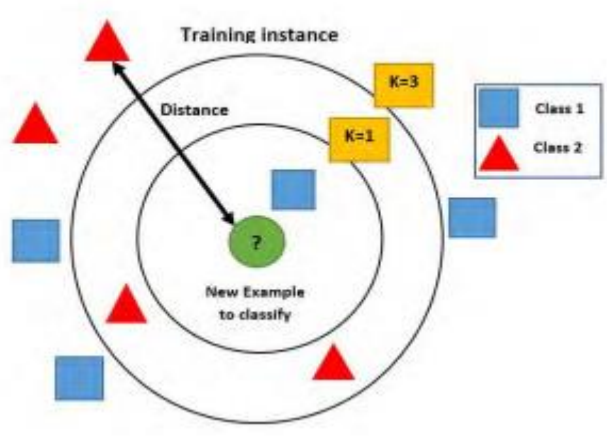

Figure 3.7: K-nearest algorithm [56]

3.5. EVALUATION METHODS

In order to choose the best model or technique from a group of available models or techniques, the evaluation is crucial for understanding the model's or technique's quality and for fine-tuning parameters in the iterative learning process [64]. Any model's effectiveness and performance can be evaluated using this method. With the k-means, filtered, and farthest first algorithms, fifteen experiments are conducted in this study. The clustering performance is validated and compared based on the number of iterations, inter-class similarity error (SME), percentage of incorrectly clustered data, and time to build the model. Inter-class similarity error, also known as SME, is a misclassification that happens when two distinct classes or categories of data are incorrectly identified as being similar or falling under the same class.

Rate of incorrectly clustered: refers to the proportion or percentage of data points that are assigned to the wrong cluster during a clustering process. It is a measure of the clustering algorithm's accuracy in correctly grouping similar data points together.

Rate of incorrectly clustered = (Number of incorrectly clustered data points) / (Total number of data points).

For classification 9 experiments conducted in J48 decision tree, KNN and Naïve Bayes to validate the performance, accuracy, recall and precision of the model are considered.

Accuracy: is a commonly used metric for evaluating the performance of classification algorithms. It measures the proportion of correctly classified instances out of the total number of instances in a dataset. In other words, accuracy tells us how often the classifier makes correct predictions.

Accuracy = (Number of correct predictions) / (Total number of instances)

Recall: is a metric used in classification algorithms to measure the ability of a model to correctly identify all positive instances from the total number of actual positive instances in a dataset. It quantifies the percentage of true positive predictions made by the classifier.

Recall = (Number of true positives) / (Number of true positives + Number of false negatives)

Precision: is a metric used in classification algorithms to measure the accuracy of positive predictions made by the model. It quantifies the percentage of correctly predicted positive instances out of the total number of instances predicted as positive.

Precision = (Number of true positives) / (Number of true positives + Number of false positives)

CHAPTER FOUR

PROBLEM UNDERSTANDING AND DATA PREPARATION

4.1 OVERVIEW

The fundamental steps followed by this research are understanding the problem domain, comprehending the data, preparing the data, data mining, evaluating the discovered knowledge, and applying the knowledge. In particular, problem comprehension, data comprehension, and data preprocessing are covered in the study's data preparation phase, which is covered in this chapter.

4.2. UNDERSTANDING OF THE PROBLEM

It is clear that customer segmentation aids in understanding the real needs of customers and their banking practices. Additionally, customer segments like value baskets, value drivers, or revenue engines should be taken into account. An organization can better understand the relative importance each customer represents in relation to overall sales and profits by segmenting customers into portfolios. Through the implementation of strong relationship development, such understanding will help businesses not only retain valuable customers but also add value to their relationships with them. Because bank services and products should be based on a better understanding of customers, customer identification aids banks in boosting profitability [9].

As the bank continues to grow its business, it should gear efforts to grow with its customers. Hence by refining offerings according to the needs of the customers, providing advisory services and by making their everyday operations more efficient, the bank will assist its customers in their journey

towards growing their business. Throughout a customer's life cycle in the bank, some clients could grow from segment to segment. This should be properly monitored and changes to the marketing segmentation, sales and value offers and service levels help leverage this growth.

Categorizing a customer in a certain segment should ultimately be reasoned out or justified by the benefits generated due to the practice being in place. In principle, the very purpose of customer portfolio management practice is to ensure sustainable customer base and value generation capability of the Bank as a going-concern. The type of segmentation used depends on the specific business objective and target. Different criteria and segmentation methods are appropriate for different situations and business objectives [5]. As below listed customer segmentation criteria, customers are segmented in to small, medium and premium customer segments based on below listed point.

1. Segmentation of Small Customer Segment
   - Identify accounts opened by individual customers or non-business entities.
   - Compute monthly credit turnover for the account.
   - Identify the customer segment where the monthly credit turnover and the disposable income range for the segmentation criteria match.

2. Segmentation of Medium Customer Segment
   - Identify accounts opened by business entities.
   - For non-credit customers, the customer segmentation is carried out to know their credit turnover of the account balance.

3. Segmentation of Premium Customer Segment
   - Identify saving accounts and checking accounts opened by business entities and/or institutions.
   - Foreign direct investment, multi-national companies, international NGOs and organizations, embassies, government agencies and export business entities should be segmented under the premium banking.

4.2.1 CUSTOMER SEGMENTATION CRITERIA

As per a discussion made with domain experts customer segmentation criteria is based on customer's initial deposit and monthly credit turnover on the customer account and derive segmentation criteria, as listed in the below table 4.1.

Market Segment	**Alternative Segmentation Criteria[ETB]**	
	Initial Deposit (Ini.D**)**	**Monthly Credit turnover**
Small customer segment	Ini.D<=100,000	<=500,000
Medium customer	100,001<=Ini.D<=200,000	>=500001 and <=1000000

segment		
Corporate customer segment	Ini.D>=200,001	>=1000001

Table 4.1Criteria for customer segmentation

4.2.2 CUSTOMER SEGMENT AND VALUE PROPOSITIONS

As the primary purpose of customer segmentation is to enhance capability of the bank in understanding the true demand of customers and customers are segmented based on similarity of needs and banking habits. The decision to divide the customer groups into three segments is primarily based on the preference of the bank's business team. They believe that managing a larger number of segments would become challenging and assigning a value proportion to each group would be difficult. Following expert recommendations, the researchers have grouped the customers into three segments and assigned a value proportion to each segment which is portrayed below in table 4.2.

Market Segment	**Value-Propositions**
Small customer segment	• Favorable saving interest rates • Amex Platinum Credit Card linked to Saving account • Special checking account linked to saving account • Consumer Financing loan • Amex Gold Card • Education loan • Medical loan • Idea financing • Event financing • Zero balance account opening
Medium customer segment	• Flexible account opening • Special savings schemes • Working capital • Lease financing • Personal guarantee facilities • Buyers and suppliers guarantee facilities
Corporate customer segment	• Special savings schemes • Favorable interest rate for long-term fixed time deposits • Working capital financing • Value Chain financing • Tender, performance and advance payment security and bid bonds • Leasing and equity financing

	• Foreign exchange • Cash handling • Payroll handling through door to door banking

Table 4.2Fundamental value proposition of the Bank for customer segment

4.2 DATA UNDERSTANDING

To make customer segmentation criteria actionable, the study defines attractive and well-differentiated offerings. Customers in each segment should have similar demand patterns. And the Implementation of the segmentation is largely carried out by the bank's core banking data. Identification of customers is done with currently available data in the bank's system.

Having an insight about the need of data mining the next basic thing for the process is getting the customers data and creating an understanding for it. Data understanding step of hybrid-DM has different components of learning before the actual application of data mining techniques. This step includes collecting data and checking for completeness, redundancy, missing values, credibility of attribute values. Finally, the step includes verification of the usefulness of the data with respect to the DM task [57].

4.2.1 INITIAL DATA COLLECTION

The data for this research has been compiled from Dashen bank Core banking system Database (CBS). The core banking system contains different tables related with customers' activities and many other modules of the bank. Among these tables' the researcher considered only a data relevant to the study, which are taken from customer account maintenance, customer transaction and customer function user defined.

- Customer account maintenance table: - this table contains customer account related information such as customer branch, customer account, current available balance of the account, customer type, customer address, account open date. The table has used to generate information related to maintained account by the bank.
- Customer function user defined table: this table maps the customer by their transaction history to the predefined categories. The information contained in this table include customer transaction reference with its related account type and other user defined fields like initial deposit.
- Customer transaction table: this table contains all account and also contract related transactions, such as account debit/credit history, contract linked account transaction, interest debit/credit, ATM transaction, Internet banking, and other banking incoming / outgoing transactions.

Accordingly, the following list of attributes (see table 4.5 below) are taken from the above mentioned three tables.

No	FIELDS	DESCRIPTION	DATA TYPE
1	Branch_Name	Branch Name of the bank	Nominal
2	Location_Code	The location of the customer	string
3	Nationality	The Nationality of the customer	string
4	Limit_Facility	Over drawn facility	Nominal
5	Customer_Category	Category of the customer based on the account purpose	Nominal
7	Account_Class	Service subscribed by the customer	Nominal
8	Account Currency	Account currency linked to the account	Nominal
9	Demand_Account	Cheque account	Nominal
10	Saving Account	Default customer account type	Nominal
11	Monthly_Credit_Tov	Credit and debit turnover of the customer account	Nominal
12	Initial Deposit	Starting deposit when the account open	Numeric

Table 4.3 List of attributes with their description

4.3 DATA PREPROCESSING

Collecting, analyzing, and preparing high-quality data for modeling with data mining algorithms is a crucial step in completing the research. In actuality, data mining cannot be performed directly on the dataset. Data preprocessing tasks must be applied in order to further clean and transform the data. Thus, this research project includes procedures for data transformation and cleansing. Preprocessing helps to fill in some missing values, identify some outliers that could compromise the outcome of data mining, and identify, remove, or correct some noisy data, according to Han and Kamber [4]. It is necessary to carry out data normalization, discretization, and related tasks in connection with this. It is also necessary to prepare the dataset in the right format in order to perform the experiment.

4.3.1 DATA SELECTION

On this phase, the relevant data for the data mining process is selected. The cornerstone of selection of data was based on relevance, availability and quality of variables.
The most challenging part of this step was getting the relevant information as needed. Most of the core banking reports is designed for daily routine and periodic administrative decisions per branch and specific products.

The core banking system, stacked with various tables and views is not an easy task to get the desired information without losing focus. Identifying the important ones and further reducing the number of tables to minimize the number of joins has taken extended time. Finally, the tables and views mentioned in the data description part were selected with the relevant fields. While selecting the data the following inclusion criteria are used:

- Active customers are selected.
- Customer accounts which has movement in their account are selected.
- PF (provident fund) accounts, cash-in-cash-out accounts are excluded.
- Only conventional accounts are selected.
- The selection of 12 branches is based on several factors including the number of customers, number of transactions, establishment date, and target customer demographics.

4.3.2 DATA CLEANING

To fill in missing values, handle outliers, smooth noisy data, find inconsistencies, and fix the data set, data cleaning uses a variety of techniques [58]. Data cleaning entails removing records from which each attribute column contained information that was inconsistent, duplicated, incomplete, or irrelevant. There are several ways to deal with missing values, including disregarding tuples, filling in the gaps with the modal value (for nominal and ordinal variables), and the mean (for continuous variables) [59]. The initial deposit field in the data taken from the core banking database used in this study only has 3472 missing values. These values are overlooked when customer accounts are opened with no deposit balance, according to our conversations with experts. Since none of the other attributes have noisy data to clean, the missed values are therefore replaced with zero values.

4.3.3DATA TRANSFORMATION AND AGGREGATION

Data transformation and aggregation assists in reducing the variations of field values and changes to a meaningful and understandable form. In this research the following attributes data type is transformed to conceptual definition of each categorical value.

- Monthly credit turnover field is the derived value by subtracting debit turnover from total credit turnover and the values are change into nominal values.
- Initial Deposit values are changed to nominal values.

Field	Value	Nominal values
Monthly_credit_turnover	<200000	SMALL

	>200000 and <1000000	MEDIUM
	>1000000	CORPORATE
Initial Deposit	<100000	BASIC
	>100000 and <5000000	MSME
	>500000	TOP

Table 4.4 Attributes data type transformation.

4.3.4 DATA FORMATTING

Data formatting is the activity of changing the data format into a format suitable or understandable by the data mining tool. The datasets should be transformed into a format that is acceptable by WEKA tool. ARFF (Attribute-Relation File Format) is a format used by the WEKA tool for running and creating a model. The data set is first saved into a CSV comma-separated format (see figure 4.6) and then saved in WEKA understandable file format called ARFF (Attribute-Relation File Format).

```
@relation 'final dataset new test-weka.filters.unsupervised.attribute.Remove-R9,11'

@attribute BRANCH_NAME {'BAHIR DAR BRANCH','AWASSA BRANCH','DESSIE BRANCH','GONDAR BRANCH','BEKLO BET  BRANCH','BOLE MEDHANIAI
@attribute CCY {ETB,USD,EUR,GBP}
@attribute CUSTOMER_CATEGORY {STAFF,INDIVIDUAL,C,I,PLC,DIASPORA,SHARECOM,GOVTMNT,WALKIN,JOINTVENT,CORP,ASSOC,BANK,NGO,PUBLICEI
@attribute LOC_CODE {AAB,CIF,UDS,ET,UAW,UDD,UGO,UMK,UAB,CN,UAD,UDL,UDW,UPC,UJM,UAL,ULA}
@attribute NATIONALITY {ET,SA,IN,US,GB,FR,AE,EU,NL,IS,CN,DJ,SE,SU,YER,ER,KE,BF}
@attribute LIMIT_FACILITY {NO,YES}
@attribute DEMAND_ACCOUNT {NO,YES}
@attribute SAVING_ACCOUNT {YES,NO}
@attribute Deposit {BASIC,MSME,TOP,RETAIL}
@attribute CUSTOMER_CLASS {SMALL,MEDIUM,CORPORATE}

@data
'BAHIR DAR BRANCH',ETB,STAFF,AAB,ET,NO,NO,YES,BASIC,SMALL
'AWASSA BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,MEDIUM
'BAHIR DAR BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'DESSIE BRANCH',ETB,INDIVIDUAL,CIF,ET,NO,NO,YES,BASIC,SMALL
'DESSIE BRANCH',ETB,INDIVIDUAL,UDS,ET,NO,NO,YES,BASIC,SMALL
'BAHIR DAR BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'GONDAR BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'BEKLO BET  BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'BOLE MEDHANIALEM BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'BAHIR DAR BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'AMOUDI BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'SARIS BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'DIRE DAWA  BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'BAHIR DAR BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'GONDAR BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'BAHIR DAR BRANCH',ETB,INDIVIDUAL,ET,ET,NO,NO,YES,BASIC,SMALL
'BAHIR DAR BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,MSME,SMALL
'AWASSA BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'KERRA BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'BAHIR DAR BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'BAHIR DAR BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
'SARIS BRANCH',ETB,INDIVIDUAL,AAB,ET,NO,NO,YES,BASIC,SMALL
```

Fig 4.6Sample data in CSV format

Once the dataset is well prepared to make sure its quality and converted to WEKA understandable file format, different experiments are performed using data mining clustering and classification

algorithms for constructing descriptive and predictive models. The experimental results, analysis and discussion are presented in the next chapter five.

CHAPTER FIVE

EXPERIMENTAL RESULTS AND DISCUSSION

5.1 OVERVIEW

In this chapter, the researcher depicts the actual application of data mining process in a stepwise fashion on the customers of Dashen Bank. As discussed with domain expert of the bank specifically innovation and development department has been doing the analysis using statistical methods on the data which is extracted from the database and also from branches. However, with the help of statistical method the domain experts are unable to use the whole data in order to reach to the desired analysis results. To solve this problem, the researcher use data mining techniques. Using these data mining techniques hidden knowledge is discovered from customer's transactional data which is used to solve customer segmentation problem. For the clustering and classification analysis of this research 13 attributes and 32674 training instances and 13024 test datasets are selected and the experiments are done using the WEKA tool.

5.2 EXPERIMENTAL SETUP

The objectives of cluster analysis are the organization of objects into groups, according to the similarity among them. In other words, clustering is unsupervised data mining techniques, which is applied when the class levels of the training data are unknown. Among the different clustering algorithms in WEKA, K-means algorithm, filtered cluster algorithm and farthest first are selected for experimentation in this study.

In this study, 15 experiments are performed with k-means, filtered and farthest first algorithms In order to validate and compare the clustering performance of clustering model done in a way that the attribute value of each cluster in the model is compared to other clustering models using parameters such as number of iterations, inter-class similarity error, time to build the model and number of incorrectly clustered. Through discussion with domain expert and advices to set k to 3 and to get better grouping results and having seed size of 10, 100, 1000, a seed size refers to the initial set of points or centroids used to start the clustering process. These initial points serve as the starting positions for the algorithm to iteratively assign data points to clusters and update the centroids until convergence is reached. initialization method Random and canopy is used to see how it changes the variable of each of clusters in these research.

Table 5.1 presents experiment conducted in the study to come up an optimal clustering model.

Experiment	Algorithms	Parameters	Initialization Mode
1	K-means	k=3, seed 10	Random
2	K-means	k=3, seed 100	Random
3	K-means	k=3, seed 1000	Random
4	K-means	k=3, seed 10	Canopy
5	K-means	k=3, seed 100	Canopy
6	K-means	k=3, seed 1000	Canopy
7	Filtered cluster	k=3, seed 10	Random
8	Filtered cluster	k=3, seed 100	Random
9	Filtered cluster	k=3, seed 1000	Random
10	Filtered cluster	k=3, seed 10	Canopy
11	Filtered cluster	k=3, seed 100	Canopy
12	Filtered cluster	k=3, seed 1000	Canopy
13	Farthest first cluster algorithm	k=3, seed 10	No
14	Farthest first cluster algorithm	k=3, seed 100	No
15	Farthest first cluster algorithm	k=3, seed 1000	No

NB: k is the number of clusters

Table 5.1 Experiments conducted with test mode

5.3 EXPERIMENTAL RESULT

Before starting the experimentation, it is important to set the threshold value for numeric attributes of the data sets. For this research purpose, thirteen attributes are selected based on data preprocessing method as discussed in chapter three, which are selected attributes to discover hidden knowledge so as to segment and identify the customer behavior.

5.3.1 CLUSTERING RESULT USING K-MEANS ALGORITHM

As stated in table 5.1, using k-means clustering algorithm six experiments conducted. Experimental result is shown in table 5.2 below.

Experiment	**Algorithm**	**Parameters**	**Initialization method**	**SSE**	**Model building time in seconds**	**Incorrectly clustered**

						(in %)
1	K-means	K=3, seed 10	Random	42226	0.08	24.09%
2	K-means	k-3, seed 100	Random	38824	0.04	22.93%
3	K-means	k-3, seed 1000	Random	38824	0.06	22.93%
4	K-means	k-3, seed 10	Canopy	47268	0.09	11.17%
5	K-means	k 3, seed 100	Canopy	46221	0.07	7.35%
6	K-means	k 3, seed 1000	Canopy	46221	0.08	7.44%

Table 5.2 Performance result for K-means algorithm

The first experiment is conducted by k value to 3 and default seed value to 10 and use Random initialization method has generated a model with a number of iteration 4, time is taken to build the model 0.08 second; within-cluster sum squared error is 42226 and 24.09% incorrectly clustered instances. The second experiment is conducted by k value to 3 and seed value to 100. The experiment has generated a model with a number of iteration 3, time taken to build the model 0.04 second; within sum squared error is 38824 and 22.93% incorrectly clustered instances. The third experiment is conducted by k value to 3 and seed value to 1000. The experiment has generated a model with a number of iteration 3, time taken to build the model 0.06 second; within sum squared error is 38824 and 22.93% incorrectly clustered instances. The fourth experiment is conducted by k value to 3 and seed value to 10 and initialization method Canopy. The experiment has generated a model with a number of iteration 2, time taken to build the model 0.09 second; within sum squared error is 47268 and 11.17% incorrectly clustered instances. The fifth experiment is conducted on k value 3 and seed size 100 the experiment has generated a model with a number of iteration 2, time taken to build the model 0.07 second; within sum squared error is 46221 and 7.35% incorrectly clustered instance. Sixth experiment is k value 3 and seed size 1000 with canopy initialization method. The experiment generated the model with a number of iteration 2, time taken to build the model is 0.08 seconds, and within sum squared error is 46221 and 7.44% incorrectly clustered instances.

Generally, from the six experiments conducted before, the model developed with the k value 3 and seed value 100 with initialization method random test option has a minimal SSE value; hence, it gives better performance of identifying the value and behavior of the customer. Therefore, among the six experiments of k-means algorithm models built in the forgoing experimentations, k = 3 and seed value 100 with Random initialization method is selected.

5.3.2 CLUSTERING RESULT USING FARTHEST FIRST ALGORITHM

The farthest first clustering algorithm is one of the clustering algorithms which is choose for this research work. The farthest first clustering algorithm has some procedures related to the K-means clustering algorithm. In this algorithm, centroids are selected and assign the objects in the clusters. Table 5.3 shows results of farthest first clustering algorithm.

Experiment	Algorithm	Parameters	SSE	Model **build**ing time in **second**s	Incorrectly clustered (in %)
1	Farthest First	k=3, seed 10	No	0.06	8.29%
2	Farthest First	k=3, seed 100	No	0.02	8.22%
3	Farthest First	k=3, seed 1000	No	0.02	8.21%

Table 5.3 Performance result for farthest first algorithm

The first experiment is conducted by k value to 3 and seed value to 10. The experiment has built a model taking time of 0.06 seconds and 8.29% incorrectly clustered instances. The second experiment is conducted by k value to 3 and seed value to 100. The experiment takes 0.02 seconds to build the model and 8.22% incorrectly clustered instances. The third experiment is conducted by k value to 3 and seed value to 1000. The third experiment has generated a model time taken to build the model 0.02 second and 8.21 %incorrectly clustered instances.

Generally, from the three experiments conducted using farthest first algorithm, the model developed with the k value 3 and seed value 1000 test option gives better performance of identifying the behavior of the customer. Therefore, among the three experiments farthest first algorithm model built in the forgoing experimentations, k value 3 and seed value 1000 is selected

5.3.3 CLUSTERING RESULT USING FILTERED CLUSTER ALGORITHM

Using filtered cluster algorithm six experiments conducted by k value to 3 and seed value to 10, 100, 1000. The experimental result is shown in table 5.4 below.

Experiment	**Algorithm**	**Parameters**	**Initialization method**	**SSE**	Model **build**ing time in **second**s	**Incorrectly clustered**
1	Filtered	K=3, seed 10	Random	40011	0.1	27.49 %
2	Filtered	K=3, seed 100	Random	39682	0.06	29.25 %
3	Filtered	K=3, seed 1000	Random	39949	0.07	24.28 %
4	Filtered	K=3, seed 10	Canopy	46312	0.1	7.35%
5	Filtered	K=3, seed 100	Canopy	46312	0.12	7.44%
6	Filtered	K=3, seed 1000	Canopy	46312	0.11	7.44%

Table 5.4 Performance result for filtered algorithm

The first experiment is conducted by k value to 3 and default seed value to 10 and initialization method Random experiment has generated a model with a number of iteration 4, time is taken to build the model 0.01 second; within-cluster sum squared error is 40011 and 27.49 % incorrectly clustered instances. The second experiment is conducted by k value to 3 and seed value to 100. The experiment has generated a model with 2 iterations, time taken to build the model 0.06 second; within sum squared error is 39683 and 29.25% incorrectly clustered instances. The third experiment is conducted by k value to 3 and seed value to 1000. The experiment has generated a model with 2 iterations, time taken to build the model 0.07 second; within sum squared error is 39949 and 24.28% of incorrectly clustered instances. The fourth experiment is conducted by k value to 3 and seed value to 10 and initialization method Canopy. The experiment has generated a model with 2 iterations, time taken to build the model 0.01 second; within sum squared error is 46312 and 7.35 % of incorrectly clustered instances. Fifth experiment is conducted on k value 3 and seed size 100 the experiment has generated a model with 3 iterations, time taken to build the model 0.12second; within sum squared error is 46312 and 7.44% of incorrectly clustered instances. Sixth experiment is k value 3 and seed size 1000 with canopy initialization method. The experiment generated the model with 2 iterations, time taken to build the model 0.11 seconds within sum squared error is 46312 and 7.44 % of incorrectly clustered instances.

Generally, from the six experiments conducted before, the model developed with the k value 3 and seed value 100 with Random initialization method option given better performance of identifying the value and behavior of the customer. Therefore, among the six experiments of filtered algorithm models built in the forgoing experimentations, k value 3 and seed value 100 is selected.

5.3.4 COMPARISON OF CLUSTERING ALGORITHMS RESULTS

As per the aim of this study to select the best clustering technique for building descriptive model for customer segmentation three clustering algorithms are experimented namely; k-means, filtered and farthest first. Then, 15 different experiments conducted and the obtained optimal results of the three algorithms are compared in table 5.5 below.

Clustering Algorithm	SSE	Model building time in seconds	Accuracy
K Means	38824	0.04	92.55%
Farthest First	No	0.02	91.78%
Filtered Cluster	39682	0.06	70.74%

Table 5.5 Performance Comparison of the clustering Algorithms

From the conducted experiments, the experiment who took less value of SSE measures the compactness, or how similar the data points within each cluster are. It is defined as the sum of the squared distances between each data point and its assigned centroid or cluster center. The objective of clustering algorithms such as K-Means is to minimize SSE, as smaller SSE indicates that the data points are more tightly clustered around their respective centroids. Hence, k-Means clustering algorithm outperformed the Filtered and Farthest First clustering algorithms for enterprise customer segmentation. The k-Means algorithm took 0.04 seconds to build the model with 92.55% accuracy, which was validated using 32674 instances from the CBS database. In comparison, Farthest First

clustering achieved an accuracy rate of 91.78% and took the second longest time to build the model, while the Filtered clustering algorithm took the longest time at 0.06 seconds and achieved an accuracy rate of 70.74%.

By comparing the results presented in Table 5.5, it is clear that the k-Means clustering algorithm with k value 3 and seed 100 has the highest accuracy and least time to build the model, making it the preferred choice for the study. Overall, the m^2 error in clustering was minimized with the k-Means algorithm, indicating its effectiveness for enterprise customer segmentation. Once an optimal descriptive model is selected, it is used to label the unlabeled data and submit it for classification algorithms for constructing a predictive model.

5.4 CLASSIFICATION MODELING

After clustering the data, the output can be saved in an. arff format and used as an input for classification in Weka. The. arff format is a standard machine-readable file that describes how to interpret the data within it. This makes it easier to use with Weka classifiers, allowing us to quickly train models on our clustered dataset and apply them for predictive tasks such as classification or regression. In this research, as explained in methodology part, to build classification model, the output of clustering model that is built through different experiments and chosen as a best model, has been used as an input for the purpose of constricting a predictive model for identifying potential customers. The algorithms selected for classification purpose are J48 decision tree, naive Bayes and KNN. The researcher tested the algorithm with different parameters and record numbers to improve the classification accuracy. Finally, models are compared and the best model is selected. The selected classification model generates rules that enable to identify profitable customers in the customer segmentation. Table 5.6 below presents experiments conducted in the study to come up with an optimal classification model.

Experiment	Algorithms	Test Mode
1	Decision Tree	Decision tree with70/30
2	Decision Tree	Decision tree with10-fold cross validation
3	Naïve Bayes	Naïve Bayes with 70/30
4	Naïve Bayes	Naïve Bayes with 10-fold cross validation
5	KNN	KNN with 70/30
6	KNN	KNN with 10-fold cross validation

Table 5.6 Experiment setup

5.4.1 DECISION TREE MODEL BUILDING

One of the most frequently employed supervised learning techniques is the decision tree classifier. The real goal of the decision tree is to categorize the data into distinct groups or branches that produce the strongest separation in the values of the dependent variable, making it superior to predict segments with a desired individual behavior, such as response or activation, and thereby offering an answer that is simple to understand [58].

Below in Table 5.7, input dataset and the result of decision tree output at different experiments are illustrated

Experiment	No. of instances	Test Mode	Confidence factor	Time taken to build model (in seconds)	Accuracy
1	32674	10-fold cross validation	C-0.45	0.08	92.00%
2	32674	10-fold cross validation	C-0.90	0.59	91.99%
3	13709	70/30	C-0.45	0.06	92.07%
4	13709	70/30	C-0.90	0.03	92.08 %

Table 5.7 Performance result of Decision tree

From all experiments conducted in this study, the use of 10-fold cross validation with confidence factor of 0.45 results in classification accuracy of 92%. The second experiment is using 10-fold cross validation with confidence factor of 0.90 results in classification accuracy of 91.99%. The third experiment is using 70/30 percentage split test option with confidence factor 0.45 and accuracy of the model is 92.07%. The final experiment conducted is 70/30 test mode with confidence factor of 0.90 results in classification accuracy is 92.08%. From all experiments carried out the researcher select 70/30 percentage split test model with confidence factor 0.90 and accuracy of classification 92.08% which has a better accuracy of result.

5.4.2 NAIVE BAYES MODEL BUILDING

In the study, Naive Bayes classifier is also tested to see its performance in Dashen bank customer's prediction. Experimental result of Naive Bayes classifiers is presented in table 5.8 below

Experiment	**No. instances**	**Test Mode**	**Time taken to build model (in seconds)**	**Accuracy**
1	32674	10-fold cross validation	0.05	90.70%
2	13709	70/30	0.04	90.56%

Table 5.8: Summary of experimental result of Naive Bayes Classifier Algorithm

By analyzing the performance of the models produced by Naive Bayes, the highest accuracy of 90.70% is achieved by 10-fold cross validation. Therefore, among the two experiments of Naïve Bayes algorithm the model created by10-fold cross validation is selected.

5.4.3 K-NEAREST NEIGHBORS MODEL BUILDING

In the study, KNN classifier is also tested to see its performance in Dashen bank customers' prediction. Experimental result of KNN classifiers is presented in table 5.9below.

Experiment	**No. instances**	**Test Model**	**Time taken to**	**Accuracy**

			build model (in seconds)	
1	44313	10-fold cross validation	0.01	91.97%
2	13709	70/30	54.82	92.04%

Table 5.9: Summary of experimental result of KNN Classifier Algorithm

By analyzing the performance of the models produced by KNN, the highest accuracy of 92.04% is achieved by 70/30 test mode. Therefore, from the conducted experiments of KNN algorithm the model created by 70/30 test mode is selected.

5.4.4 COMPARISON OF THE EXPERIMENTED CLASSIFICATION MODELS

A comparison of J48 decision tree, Naïve Bayes and KNN Classification models is presented in the below table 5.10.

Parameters	J48 **decision tree** with 70/30 test mode	Naïve Bayes with 10-fold cross validation	KNN with 70/30 test mode
Average TP Rate	0.921	0.907	0.920
Average FP Rate	0.698	0.587	0.920
Average Precision	0.892	0.891	0.891
Average Recall	0.921	0.907	0.920
Average F-Measure	0.900	0.898	0.899
MCC	0.373	0.375	0.372
ROC Area	0.806	0.826	0.817
PRC Area	0.908	0.912	0.909
Accuracy	**92.08%**	**90.70%**	**92.04%**

Table 5.10 Performance Comparison for the selected Algorithms

Among the experimented classification algorithms, J48 decision tree registered the highest accuracy of 92.09%. Accordingly, this algorithm is selected for segmenting the banks customer based on their monthly credit turnover amount.

The confusion matrix of the selected algorithm is presented as follows in table 5.11.

```
=== Confusion Matrix ===

     a     b     c   <-- classified as
 12464    54    55 |     a = SMALL
   569    81    54 |     b = MEDIUM
   295    58    79 |     c = CORPORATE
```

Table 5.11 Confusion matrix of J48 decision tree

The entries in the confusion matrix have the following meaning:

- 12464 is the number of correct predictions that an instance is small transactional customers. However, 54 and 55 instances of small transactional customers are incorrectly predicted as medium and corporate customers.
- 81 number of medium customers are correctly predicted. On the other hand, 569 and 54 instances of medium customers are incorrectly classified as small transactional and corporate customers.
- 79 instances are correctly predicted that an instance is corporate customer. On the other hand, 295 and 79 instances of corporate customers are incorrectly classified as small transactional and medium customers

5.5 EXTRACTING INTERESTING RULES

The experiments conducted using the J48 decision tree technique showed better performance as compared to KNN and Naive Bayes. Hence this algorithm is selected for generating rules. The set of rules are extracted simply by traversing through the output of the decision tree. One of the most captivating aspects of the following set of classification rules is the fact that they introduce entirely new insights and knowledge to domain experts. These rules offer a fresh perspective that has not been previously explored within the domain, making them highly intriguing and valuable. By unveiling previously unknown patterns and relationships, these rules have the potential to revolutionize the understanding of the subject matter. Experts in the field will undoubtedly find these rules to be a source of great interest and excitement, as they provide a unique opportunity to expand their knowledge and challenge existing assumptions. The novelty and potential impact of these rules make them an exciting area for further investigation and exploration. The following are the selected set of rules which are in line with the survey of customer segmentation which gain the attention of domain experts.

RULE# 1:

DEMAND_ACCOUNT = NO, CCY = ETB, Deposit = BASIC: SMALL (41390.0/2009.0)

- If the account is not a demand account, the currency is ETB, and the deposit is categorized as BASIC, then classify it as SMALL with a support of 41390.0 and a confidence of 2009.0.

RULE#2

DEMAND_ACCOUNT = NO, CCY = ETB, Deposit = MSME, LIMIT_FACILITY = NO: SMALL (1065.0/192.0)

- If the account is not a demand account, the currency is ETB, the deposit is categorized as MSME, and there is no limit facility, then classify it as SMALL with a support of 1065.0 and a confidence of 192.0

RULE#3

DEMAND_ACCOUNT = YES, Deposit = BASIC, CCY = ETB, LIMIT_FACILITY = NO: SMALL (1116.0/406.0)

- If the account is a demand account, the deposit is categorized as basic, the currency is ETB, and there is no limit facility, then classify it as SMALL with a support of 1116.0 and a confidence of 406.0

RULE#4

DEMAND_ACCOUNT = YES, Deposit = BASIC, CCY = ETB, LIMIT_FACILITY = YES: CORPORATE (16.0/7.0)

- If the account is a demand account, the deposit is categorized as basic, the currency is ETB, and there is a limit facility, then classify it as CORPORATE with a support of 16.0 and a confidence of 7.0.

RULE#5

DEMAND_ACCOUNT = YES, Deposit = BASIC, CCY = USD: CORPORATE (135.0/83.0)

- If the account is a demand account, the deposit is categorized as basic, and the currency is USD, then classify it as CORPORATE with a support of 135.0 and a confidence of 83.0.

RULE#6

DEMAND_ACCOUNT = YES, Deposit = MSME, LIMIT_FACILITY = NO, CCY = ETB: SMALL (208.0/79.0)

- If the account is a demand account, the deposit is categorized as MSME, there is no limit facility, and the currency is ETB, then classify it as SMALL with a support of 208.0 and a confidence of 79.0

RULE#7

DEMAND_ACCOUNT = YES, Deposit = MSME, LIMIT_FACILITY = NO, CCY = USD: MEDIUM (13.0/7.0)

- If the account is a demand account, the deposit is categorized as MSME, there is no limit facility, and the currency is USD, then classify it as MEDIUM with a support of 13.0 and a confidence of 7.0.

RULE#8

DEMAND_ACCOUNT = YES, Deposit = MSME, LIMIT_FACILITY = YES: MEDIUM (8.0/3.0)

- If the account is a demand account, the deposit is categorized as MSME, and there is a limit facility, then classify it as MEDIUM with a support of 8.0 and a confidence of 3.0

RULE#9

DEMAND_ACCOUNT = YES, Deposit = TOP, LIMIT_FACILITY = NO, CCY = ETB: SMALL (540.0/332.0)

- If the account is a demand account, the deposit is categorized as TOP, there is no limit facility, and the currency is ETB, then classify it as SMALL with a support of 540.0 and a confidence of 332.0

RULE#10

DEMAND_ACCOUNT = YES, Deposit = TOP, LIMIT_FACILITY = NO, CCY = USD: CORPORATE (59.0/30.0)

- If the account is a demand account, the deposit is categorized as TOP, there is no limit facility, and the currency is USD, then classify it as CORPORATE with a support of 59.0 and a confidence of 30.0

RULE#11

DEMAND_ACCOUNT = YES, Deposit = TOP, LIMIT_FACILITY = NO, CCY = EUR: SMALL (7.0/2.0)

- If the account is a demand account, the deposit is categorized as TOP, there is no limit facility, and the currency is EUR, then classify it as SMALL with a support of 7.0 and a confidence of 2.0

Summary:

The provided rules outline the classification criteria for different types of accounts based on their attributes. If an account is not a demand account, has the currency ETB, and the deposit is categorized as BASIC or MSME without a limit facility, it is classified as SMALL with varying levels of support and confidence. Similarly, if the account is a demand account, has the currency ETB, and the deposit is categorized as basic without a limit facility, it is also classified as SMALL with a higher support and confidence. However, if the account has a limit facility, it is classified as CORPORATE. Furthermore, if the account has a currency of USD and is categorized as basic, it is classified as CORPORATE with a higher support and confidence. Additionally, if the account is a demand account, has the currency ETB, and the deposit is categorized as MSME without a limit facility, it is classified as SMALL with a moderate level of support and confidence. On the other hand, if the currency is USD, it is classified as MEDIUM.

5.6 DISCUSSION OF RESULT

The researchers aimed to identify appropriate clusters for customer segmentation based on the similarity of instances. They successfully achieved this by utilizing the collected data and implementing a clustering algorithm. Furthermore, they built a predictive model that aids in classifying and determining segments of customers. This model was developed using the prepared data and the identified clusters.

To evaluate the effectiveness of the proposed predictive model, the researchers conducted performance evaluations. This allowed them to assess how well the model performed in its classification and segmentation tasks. In comparing the results of this study with previous studies, it is important to note that the focus of this research was on customer segmentation using a

predictive model. While previous studies may have touched on similar topics, this particular study offers a unique approach and methodology.

Overall, the research questions in this study have been effectively addressed through comprehensive literature review, data collection, data preparation, cluster identification, model building, and performance evaluation. The findings of this study contribute to the existing body of knowledge in the field and provide valuable insights for further research and practical applications.

Clustering and then classifying Dashen bank customers can yield a number of results. These include the following:

1. Identification of different customer segments with distinct needs and preferences. Through clustering, bankers can identify sub-groups within their customer base that have similar characteristics such as age group, income level or spending habits. This allows to create targeted marketing strategies specifically for these groups.
2. Improved targeting techniques when launching new products or services By understanding which type of customers belong to each segment, banks are better able to accurately target those most likely to be interested in a particular product or service launch and maximize its chances of success.
3. Enhance opportunities to understand customer segment that helps Dashen bank leverage existing relationships by offering additional services relevant to. It allows you to focus promotional efforts on those more likely benefit from what's being offered.

CHAPTER SIX

CONCLUSION AND RECOMMENDATION

6.1. CONCLUSION

The growth of interest in data, information and knowledge management has been helping many organizations to digitize and manage their information resource for effective use in future to prediction about their business processes, product and behavior of their customers.

Implementation of data mining technologies in Dashen Bank help to discover pattern for customer's segmentation and classification in order to enhance service delivery and to maximize profit of the bank. Data mining application could be used to cluster and classify customers‟ transaction

behavior. Then, the data mining technology could help to predict potential customers‟ from the other in taking measures to improve the service delivery and profit generating in the future. To uncover the hidden knowledge within the dataset of the bank, preprocessing, of the dataset were performed using WEKA tool. The data was analyzed and interpreted using the WEKA 3.8.5 version software. Clustering and then classification models were built to categorize and predict customers. To cluster instances into similar groups, Filtered clustering algorithms were employed. Thus, using Filtered clustering algorithm, different experiments conducted with different K-values and seed sizes. Segmentation at K=3 and seed size 100 with 3 clusters selected as best customers segment model in the bank. To classify instances into same groups based on result obtained through clustering, Decision Tree J48 algorithm was employed. Using Decision Trees J48algorithm also different experiments conducted. Model built 70/30 test mode which registered high accuracy (92.08%), selected as best model for prediction purpose.

The finding's result demonstrates that there are patterns relationships among different attributes. Thus, based on discovered patterns, customers can be clustered in different groups according to their similarities. Then, it will be easy to predict and classify customers accordingly. Experimental findings show that there are different options to attract and retain customers, as well as to gain competitive advantage in the industry and also to serve customers with optimal satisfaction based on the discovered patterns. These paternal relationships between the data indicate the hidden fact among the dataset. Therefore, the bank could plan and implement strategies for effective and efficient service to maximize profit gained through maximum customer satisfaction.

Generally, to predict the behavior of customers, and business processes of the bank, data mining techniques offers great promise in helping the bank to discover hidden patterns in their data. Thus, using combination of techniques likes clustering with classification (as done in this research work) will assist to overcome the complexity of problems in business processes. In this research work customers are analyzed and segmented on their transaction balance, in which we obtained promising result. The limitations of this study are twofold. Firstly, the analysis focuses solely on conventional customers, while Islamic Financial Banking (IFB) customers are not included in the study. This omission is due to the operational differences between conventional and IFB customers, which may impact their behaviors and preferences. Therefore, the findings of this study may not be representative of the entire customer population of the bank.

Secondly, the study relies exclusively on transactional data to segment customers. While transaction data can provide valuable insights into customer behavior, it may not capture the full spectrum of customer attributes and characteristics.

This research successfully addresses the three research questions by providing the following insights and findings. Which attributes are suitable for applying data mining in customer segmentation and prediction? From the analysis, it has been found that the most effective attributes for this purpose are behavioral attributes, psychographic attributes, and customer lifetime value.

Behavioral attributes involve analyzing customer behavior, such as transaction history, frequency of transactions, product usage patterns, website interactions, and response to marketing campaigns.

These attributes can provide valuable insights into customer engagement and preferences, enabling the identification of different customer segments.

Psychographic attributes capture customers' attitudes, values, interests, and lifestyles. This includes personality traits, values, hobbies, opinions, and social media behavior. Psychographic attributes play a crucial role in identifying segments based on shared motivations and preferences.

Customer lifetime value (CLV) is a metric that estimates the total value a customer brings to a business over their entire relationship. CLV helps identify high-value customer segments and allows for prioritization of marketing efforts accordingly.

Based on the conducted experiment, the algorithms utilized for both clustering and classification purposes have demonstrated their suitability for this study.

The proposed model shows a significant capability in identifying customer segments. By utilizing data mining techniques and algorithms, the model can effectively analyze customer attributes, behaviors, and preferences to identify distinct segments. The inclusion of suitable attributes such as behavioral attributes, psychographic attributes, and customer lifetime value enhances the accuracy and effectiveness of the segmentation. Through the analysis of customer data, the model can uncover patterns, similarities, and differences among customers, enabling the identification of meaningful segments. This segmentation allows businesses to tailor their products, services, and marketing efforts to better meet the specific needs and preferences of each segment. By doing so, businesses can enhance customer satisfaction, improve marketing effectiveness, and drive overall profitability.

Overall, the proposed model demonstrates a strong potential in identifying customer segments, providing businesses with valuable insights to better understand their customers and tailor their strategies accordingly.

6.2 RECOMMENDATION

Based on the findings of the study discussed above, the following recommendations are forwarded:

- It is recommended that Dashen Banks focus on developing an integrated data warehouse database, which is specifically designed for query and analysis, rather than for transaction processing. This type of database would enable the bank to access historical data, such as customer banking behavior, easily and efficiently through a single source. Moreover, this type of database would make it possible to segregate the workload associated with analyzing the data, from the workload associated with transactional processes. This separation will help the bank to more effectively manage their resources. It will also enable the bank to consolidate information from multiple sources and obtain detailed insights into customers' banking habits and preferences. Integrating data from various sources and consolidating it will enable the bank to gain valuable insights into customer's banking trends in terms of

their banking habits and preferences. These insights will help the bank to better serve their clients and increase overall business efficiency. In short, Dashen Banks can benefit greatly by investing in an integrated data warehouse database for query and analysis purposes, which will not only provide them with quick and easy access to historical data but also help them to make informed decisions and gain insights into their customers' banking behavior.

- Based on the results of its research, it has been recommended that Dashen Bank adopt and implement a comprehensive Customer Relationship Management (CRM) system. A CRM system is a powerful tool used by organizations to effectively manage customer interactions and data. The system allows businesses to effectively collect, organize, analyze, and utilize customer information in order to better understand the needs and preferences of their customers, create personalized experiences, and build stronger relationships with existing and potential clients. By adopting a CRM system, Dashen Bank would be able to maintain regular and consistent contact with its current and potential customers and leverage their data to develop more effective and targeted marketing campaigns. The benefits of such a system include improved communication and collaboration across departments, automated processes for loan collections, increased efficiency through centralized data storage, and the ability to streamline marketing campaigns with targeted messaging based on customer profiles. The CRM system would also enable the bank to generate in-depth reports that can track performance, predict future market trends, and improve forecasting accuracy. Implementing a CRM system would also lead to significant cost savings for Dashen Bank, as it would help to reduce the time and costs associated with manual market segmentation and other manual processes. Overall, a robust CRM system would provide Dashen Bank with a powerful tool to improve customer service, enhance marketing efforts, and drive overall business performance.

- Developing customer profiles is an essential element of successful marketing strategy. Companies use various criteria to categorize their customers and comprehend their needs, behaviors, and preferences. This practice helps to target the right demographics and improve sales and customer loyalty. There are two ways to create customer profiles - by collecting data from the existing database or by analyzing current banking patterns and preferences. By doing so, banks can recognize customers' likes and interests concerning different products and services. This information is crucial in developing consumer profiles that contain a range of factors such as age, gender, lifestyle preferences, annual income, and psychographic traits such as values and beliefs. Knowing customers' profiles can help provide an in-depth understanding of who they are, what they like, and what they expect. This understanding assists organizations in defining their target market and developing effective marketing campaigns. For instance, if a bank identifies that a significant portion of its customers belongs to a particular age group or income range, the company can customize its products, marketing messages, and platforms to target that specific group, resonating with their preferences and needs. Furthermore, customer profiling can help identify trends among different demographics, which informs managers of the best strategies to improve profit margins and enhance customer loyalty. By personalizing offerings based on individual profiles, customers feel valued and heard, ultimately leading to a better customer experience. In conclusion, developing customer profiles is crucial in establishing a deep

understanding of customer needs and preferences. With data and analysis, companies can successfully target specific demographics and develop campaigns and products that resonate with their customer base. Ultimately, this practice leads to a more lucrative business and better customer satisfaction and loyalty.

- The bank can derive benefits from utilizing this study in the following areas:
  1. Personalized Marketing: By understanding the characteristics and behaviors of different customer segments, the bank can create personalized marketing campaigns. This can include customized offers, recommendations, and targeted messaging to effectively engage customers and drive conversions.
  2. Product Development: The findings can provide insights into customer preferences, allowing the bank to develop new products or enhance existing ones to better align with customer needs. This can help the bank stay competitive and attract new customers.
  3. Risk Assessment: The identified customer segments can assist the bank in assessing and managing risks associated with different customer groups. By understanding the risk profiles of each segment, the bank can implement appropriate risk mitigation strategies and pricing models.
  4. Customer Service: The findings can inform the bank's customer service strategies. By understanding the unique needs and preferences of different customer segments, the bank can provide tailored customer support, personalized recommendations, and targeted assistance to enhance the overall customer experience.

6.3. THE WAY FORWARD

In this study, the focus is solely on analyzing conventional customers, while excluding customers of Islamic Financial Banking (IFB) due to operational differences in the business and the unavailability of a domestic banking manual for reference and understanding. However, it is suggested that future research can conduct specifically within the IFB wing of the bank, allowing for a more comprehensive analysis of customer behavior and preferences in that segment.

Furthermore, it's important to note that this study is primarily based on transaction data. However, in future research endeavors, it would be beneficial for researchers to consider segmenting the customer base based on additional values and variables beyond just transactional data. By incorporating other customer attributes such as demographics, preferences, and behavior, a more nuanced understanding of customer segments can be achieved, leading to more targeted and effective strategies for the bank.

Based on the results obtained in this study, future researchers can go for designing a knowledge based system that can provide an advice for simplifying predicting customers with their characteristics so as to adjust the bank service provided to them.

Finally, there is a need to apply data mining for other financial institutions towards customer segmentation and prediction so as to improve their customer relationship management.

REFERENCES

[1] Jayasree. (2013). A Review on Data Mining in Banking Sector. American Journal of Applied Sciences, 10(10), 1160–1165.

[2] Kazi Imran Moin, Dr. Qazi Baseer Ahmed(2012) "Use of Data Mining in Banking" International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 2,Mar-Apr 2012, pp.738-742

[3] Li, Yihao and Beaubouef, Theresa."Data Mining: Concepts, Background and Methods of Integrating Uncertainty in Data Mining" CCSC SC Stud. EJ 3(2010):2-7

[4] Fayyad, U., Piatesky -Shapiro, G., and Smyth, P. (1996). From Data Mining To Knowledge Discovery in Databases, AAAI Press / the MIT Press, Massachusetts Institute of Technology. ISBN 0-26256097-6 MIT.

[5] Hasan Ziafat, Majid Shakeri (2014)"Using Data Mining Techniques in Customer Segmentation" International Journal of Engineering Research and Applications ISSN: 2248-9622, Vol. 4, Issue 9(Version 3), September 2014, pp.70-79

[6]Das, S., & Nayak, J. (2022). Customer segmentation via data mining techniques: State-of-the-art review. Computational Intelligence in Data Mining 489-507.

[7] Mihova, V., & Pavlov, V. (2018). A customer segmentation approach in commercial banks. AIP Conference Proceedings.

[8] K. Tsiptsis and A. Chorianopoulos,(2009) "Data Mining Techniques in CRM: Inside Customer Segmentation", John Wiley and Sons, Ltd.

[9] Mohammad Aghaei (2021)"Market Segmentation in the Banking Industry Based on Customers' Expected Benefits: A Study of Shahr Bank "Iranian Journal of Management Studies (IJMS) 2021, 14(3): 629-648

[10] Koksal, Mustafa. (2013). A Comprehensive Research Design for Experimental Studies in Science Education. Elementary Education Online. 12. 628-634.

[11] K. J. Cios, W. Pedrycz, W. Swiniarski and L. A. Kurgan Data Mining: A KnowledgeDiscovery Approach New York: Springer Science Business Media, LLC, 2007

[12] Jiawei Han, Micheline Kamber and Jian Pei. (2012): Data Mining. Concepts and Techniques, 3rd Edition

[13]Shivali, Joni Birla, Gurpreet, 2015, Knowledge Discovery in Data-Mining, International Research Journal of Engineering Research and Technology (IJERT) NCETEMS – 2015 (Volume 3 – Issue 10)

[14] C. Priyadharsini, Dr. C.Antony, an Overview of Knowledge Discovery Database and Data mining Techniques, International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)-2014(Vol.2, Special Issue 1)

[15] Jogannagari, M., & Manchala, M., 2020. Data Mining: Techniques, Tools and its Challenges. International Research Journal of creative Research thought (IRJCRT) Volume 8, Issue 7 July 2020

[16] Slimani, T., & Lazzez, A. (2014). Efficient Analysis of Pattern and Association Rule Mining Approaches. International Journal of Information Technology and Computer Science, 6(3), 70–81.

[17] S.A.R. NIHA, 2017, Study of Data Mining Methods and its Application, International Research Journal of Engineering and Technology (IRJET) -2017 (Volume: 04 Issue: 11)

[18] Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research, 12(1), 217–222.

[19] S. S. Kaptan, N S Chobey(2002), "Indian Banking in Electronic Era", Sarup and Sons..

[20] Vivek Bhambri(2011) "Application of Data Mining in Banking Sector", International Journal of Computer Science and Technology Vol. 2, Issue 2

[21] Pascu, Adrian Ionut. (2018). "Data mining. Concepts and applications in banking sector" PhD, university of Craiova, Doctoral School of Economics sciences. Issue 1

[22] McDonald, M. & Dunbar, I. (2012). Market Segmentation, How to do it and how to profit from it. A John Wiley & Sons, Ltd, Publication.

[23] Juni Nurma Sari , Ridi Ferdiana, Lukito Nugroho, Paulus Insap Santosa (2016) Review on Customer Segmentation Technique on Ecommerce, Journal of Computational and Theoretical Nanoscience ·

[24] Batt, R. (2000). Strategic segmentation in front-line services: matching customers, employees and human resource systems. The International Journal of Human Resource Management, 11:3, 540-561.

[25] Bottcher, M., Spott, M., Nauck, D. & Kruse, R (2009). Mining changing customer segments in dynamic markets, Expert Systems with Applications 36 (2009) 155–164.

[26] Buttle, F. (2009). Customer Relationship Management: Concepts and Technologies.

[27] Lambert, M. (1990). Segmentation of Markets Based on Customer Service, International Journal of Physical Distribution & Logistics Management, Vol. 20 Iss 7 pp. 19 – 27.

[28] Fang E., Palmatier R.W., & Steenkamp, J.B.E.M. (2008). Effect of service transition strategies on firm value. J Mark, 72(4), 1–14

[29] Pauline, R. (2009). Successful Customer Service: Get Brilliant Results Fast. Crimson Publishing, Limited, Richmond;Plymouth.

[30] Anderson, J.C., Narus, J. A. and Narayandas, D. (2009). Business Market Management Understanding, Creating and Delivering Value, Pearson Education Prentice Hall: Upper Saddle River New Jersey, Third edition.

[31] Ali Serhan Koyuncugil and Nermin Ozgulbas Baskent University, Turkey,(2011),Surveillance Technologies and Early Warning Systems: Data Mining Applications for Risk Detection, Published In the United States of America by Information Science Reference (an imprint of IGI Global)

[32] A.Begunca, (2017) "soft drinks consumer segmentation using benefit sought variables-case study Kosovo market," pp. 168-173.

[33] Belachew Reganie (2013) Application of Data mining Techniques for customer segmentation and prediction: the case of Buusaa Gonofa Microfinance Institution A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science in information science Addis Ababa University .Addis Ababa, Ethiopia.

[34] Belete Biazen Bezabeh,(2011) "Knowledge Discovery for Effective Customer Segmentation: The Case of Ethiopian Revenue and Customs Authority," MSc Thesis, Addis Ababa University, Addis Ababa, Ethiopia.

[35] Fikrealem. Bayissa,(2019) “Telecom Customer Segmentation using Data mining Techniques” MSc Thesis, St. Mary’s University, Addis Ababa, Ethiopia.

[36] Asuni, Joy & Koshiya, Kinjalben. (2022). Application of Data Mining Techniques in the Banking Sector. [37] Berkhin, P. (2012). A survey of clustering data mining techniques; Grouping Multidimensional Data, pp. 25–71.

[38] Dunham, M. H Stamatis (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.

[39] P. Dhandayudam, 2012 "an improved clustering algorithm for customer segmentation," Journal of Engineering Science and Technology, vol. 4, no. 2, pp. 695-702.

[41]. Dunham, M. H Stamatis (2003). Data mining introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, Inc.

[42]. Hemashree Kilari*1, Sailesh Edara2, Guna Ratna Sai Yarra3, Dileep Varma Gadhiraju4 “Customer Segmentation using K-Means Clustering” International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 11 Issue 03, March-2022.

[43] Ahmed, S. T. et al. (2020) “A generalized study on data mining and clustering algorithms,” in New Trends in Computational Vision and Bio-inspired Computing. Cham: Springer International Publishing, pp. 1121–1129.

[44] K. Ravichandra Rao (2003). Data Mining and Clustering Techniques. Documentation Research and Training Center, Indian Statistical Institute, Bangalore.

[45]. Berry M. J. A. and Linoff, G. S (2000). Mastering data mining. New York, Wiley

[46]. M. Ramageri "Data mining techniques and applications" Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305

[47] .Wei-Yin, Loh (2011). Classification and Regression Trees. Journal of WIREs Data Mining and Knowledge Discovery, Vol. 1.

[48] Radhwan H. A. Alsagheer1, Abbas F. H. Alharan2, and Ali S. A. Al-Haboobi3 (2017) "Popular Decision Tree Algorithms of Data Mining Techniques: A Review" International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 6, Issue. 6, pg.133 – 142

[49] Wibawa, A. P., Kurniawan, A., Murti, D. M. P., Adiperkasa, R. P., Putra, S. M., Kurniawan, S. A., & Nugraha, Y. R. (2019). Naïve Bayes Classifier for Journal Quartile Classification. International Journal of Recent Contributions from Engineering, Science & IT

[50]. T. R. Patil, (2013) "Performance Analysis of Naive Bayes and J48 Classification Algorithm forData Classification," Int. J. Comput. Sci. Appl. ISSN 0974-1011, vol. 6, no. 2, pp. 256–261

[51] G. Dimitoglou, J. a Adams, and C. M. Jim,(2012) "Comparison of the C4.5 and a Naive BayesClassifier for the Prediction of Lung Cancer Survivability," J. Neural Comput., vol. 4, no.8, pp. 1–9

[52]. Velmurugan, T., & Santhanam, T. (2010). Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points. Journal of Computer Science, 6(3), 363–368

[53]. Nguyen, L. H. (2014). User Model Clustering. Journal of Data Analysis and Information Processing, 02(02), 41–48

[54] . Lubis, Z., Sihombing, P., & Mawengkang, H. (2020). Optimization of K Value at the K-NN algorithm in clustering using the expectation maximization algorithm. IOP Conference Series, 725(1), 012133.

[55]. Karim, M. M., & Rahman, R. M. (2013). Decision Tree and Naive Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing. Journal of Software Engineering and Applications, 06(04), 196–206

[56]. Chowdhury, S., & Schoen, M. P. (2020). Research Paper Classification using Supervised Machine Learning Techniques. In 2020 Intermountain Engineering, Technology and Computing (IETC)

[57] Aliyev, M., Ahmadov, E., Gadirli, H., & Alasgarov, E. (2020). Segmenting Bank Customers via RFM Model and Unsupervised Machine Learning.

[58] Rana Soudagar, "Customer Segmentation and Strategy definition in segments: in case of an internet service provider in Iran," internet provide in Iran, 2007. [102] D.T.Larose, Discovering

[59] P. Chapman,J. Clinton ,T.Khabaza,T.Reinartz and Rüdiger Wirth , "The CRISP-DM Process Model," 1999.

[60] Jiban Kpal (2011) Usefulness application of data mining in extracting information from different perspective Annals of Library and Information Studies Vol. 58.Pp7-16

[61] Begam, S. (2021b). Customer Profiling and Segmentation – An Analytical Approach To Business Strategy In Retail Banking.ustomer-profiling-and-segmentation-an-analytical-approach-to-business-strategy-in-retail-banking/

[62] J.L.Bentley ,1999 "Multidimensional Binary Search Trees Used For Associative Searching," Association for Computing Machinery, vol. 18, no. 9, pp. 509-517

[63] N.S. Netanyahu, 2002"An Efficient k-Means Clustering Algorithm: Analysis and Implementation," Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp.

[64] G.Sehgal and K.Garg, 2014 "Comparison of Various Clustering Algorithm," international journal of computer science and Iinformation Tecnology, vol. 5, no. 3, pp. 3074-3076

[65] S.M.Kumar, 2013 "An Optimize Farthest First Clustering Algorithm," in Nirma University International Conference on Engnering, Bhiwani.

[66] R.M. Shah,M. A. Butt and M. Z. Baba, 2017 "Predictive Analytic Modeling: A Walkthrough," International Journals of Advanced Research in computer Science and and Software Engineering, vol. 7, no. 6, pp. 424-426.

-[68] Tenkir S. (2009). Customs Management System in Ethiopia. (Master's Thesis), Department of Accounting and Finance, Addis Ababa University, Addis Ababa, Ethiopia.

APPENDICES

Appendix A: Clustering Algorithm

Simple K-means cluster algorithm K-3 ,Seed-100 initialization method Random

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 38824.0

Initial starting points (random):

Cluster 0: 'KERRA BRANCH',ETB,INDIVIDUAL,ET,NO,NO,YES,BASIC
Cluster 1: 'AMOUDI BRANCH',ETB,PLC,ET,NO,YES,NO,BASIC
Cluster 2: 'BAHIR DAR BRANCH',ETB,INDIVIDUAL,ET,NO,NO,YES,BASIC

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                        Cluster#
Attribute                                Full Data             0                    1                    2
                                         (45698.0)      (37048.0)            (2221.0)             (6429.0)
=============================================================================================================================
BRANCH_NAME                          AWASSA BRANCH  AWASSA BRANCH WELLO SEFER PREMIER BRANCH  BAHIR DAR BRANCH
CCY                                            ETB            ETB                      ETB                ETB
CUSTOMER_CATEGORY                      INDIVIDUAL     INDIVIDUAL               INDIVIDUAL         INDIVIDUAL
NATIONALITY                                    ET             ET                       ET                 ET
LIMIT_FACILITY                                 NO             NO                       NO                 NO
DEMAND_ACCOUNT                                 NO             NO                      YES                 NO
SAVING_ACCOUNT                                YES            YES                       NO                YES
Deposit                                     BASIC          BASIC                    BASIC              BASIC


Time taken to build model (full training data) : 0.04 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      37048 ( 81%)
1       2221 (  5%)
2       6429 ( 14%)


Class attribute: CUSTOMER_CLASS
Classes to Clusters:

     0     1     2  <-- assigned to cluster
 34556  1165  6193 | SMALL
  1515   598   203 | MEDIUM
   977   458    33 | CORPORATE

Cluster 0 <-- SMALL
Cluster 1 <-- CORPORATE
Cluster 2 <-- MEDIUM

Incorrectly clustered instances :       10481.0  22.9354 %
```

APPENDIX B: Classification Algorithm

== Run information ===

Scheme: weka.classifiers.trees.J48 -S -C 0.9 -M 2

Relation: final dataset new training-weka.filters.unsupervised.attribute.Remove-R4-5,9,11_clustered-weka.filters.unsupervised.attribute.Remove-R1,10-weka.filters.unsupervised.attribute.Remove-R3

Instances: 32674

Attributes: 7

BRANCH_NAME

CCY

LIMIT_FACILITY

DEMAND_ACCOUNT

SAVING_ACCOUNT

Deposit

CUSTOMER_CLASS

Test mode: split 70.0% train, remainder test

=== Classifier model (full training set) ===

J48 pruned tree

DEMAND_ACCOUNT = NO

| BRANCH_NAME = BAHIR DAR BRANCH

| | Deposit = BASIC

| | | CCY = ETB: SMALL (6062.0/153.0)

| | | CCY = USD: MEDIUM (4.0/2.0)

| | | CCY = EUR: SMALL (2.0)

| | | CCY = GBP: SMALL (0.0)

| | Deposit = MSME: SMALL (213.0/28.0)

| | Deposit = TOP: SMALL (148.0/52.0)

| BRANCH_NAME = AWASSA BRANCH: SMALL (6682.0/178.0)

| BRANCH_NAME = DESSIE BRANCH

| | Deposit = BASIC

| | | LIMIT_FACILITY = NO: SMALL (3443.0/82.0)

| | | LIMIT_FACILITY = YES: MEDIUM (2.0)

| | Deposit = MSME: SMALL (89.0/12.0)

| | Deposit = TOP: MEDIUM (19.0/9.0)

| BRANCH_NAME = GONDAR BRANCH: SMALL (4018.0/97.0)

| BRANCH_NAME = BEKLO BET BRANCH

| | Deposit = BASIC

| | | CCY = ETB: SMALL (4660.0/234.0)

| | | CCY = USD: MEDIUM (7.0/3.0)

| | | CCY = EUR: SMALL (1.0)

| | | CCY = GBP: SMALL (0.0)

| | Deposit = MSME: SMALL (87.0/13.0)

| | Deposit = TOP: SMALL (72.0/41.0)

| BRANCH_NAME = BOLE MEDHANIALEM BRANCH: SMALL (2207.0/159.0)

| BRANCH_NAME = AMOUDI BRANCH

| | Deposit = BASIC: SMALL (2877.0/323.0)

| | Deposit = MSME: SMALL (59.0/29.0)

| | Deposit = TOP: CORPORATE (31.0/16.0)

| BRANCH_NAME = SARIS BRANCH: SMALL (3594.0/133.0)

| BRANCH_NAME = DIRE DAWA BRANCH: SMALL (2201.0/67.0)

| BRANCH_NAME = KERRA BRANCH: SMALL (4179.0/115.0)

| BRANCH_NAME = WELLO SEFER PREMIER BRANCH

| | CCY = ETB

| | | Deposit = BASIC: SMALL (2389.0/666.0)

| | | Deposit = MSME: SMALL (73.0/42.0)

| | | Deposit = TOP: CORPORATE (61.0/36.0)

| | CCY = USD: MEDIUM (280.0/153.0)

| | CCY = EUR: SMALL (7.0/2.0)

| | CCY = GBP: SMALL (2.0/1.0)

| BRANCH_NAME = MEKELE BRANCH: CORPORATE (3.0)

| BRANCH_NAME = TANA BRANCH: CORPORATE (4.0)

| BRANCH_NAME = ADIGRAT BRANCH: CORPORATE (1.0)

DEMAND_ACCOUNT = YES

| BRANCH_NAME = BAHIR DAR BRANCH

| | LIMIT_FACILITY = NO: SMALL (211.0/54.0)

| | LIMIT_FACILITY = YES

| | | Deposit = BASIC: SMALL (1.0)

| | | Deposit = MSME: SMALL (3.0/1.0)

| | | Deposit = TOP: MEDIUM (9.0/1.0)

| BRANCH_NAME = AWASSA BRANCH

| | Deposit = BASIC: SMALL (147.0/22.0)

| | Deposit = MSME: SMALL (18.0/6.0)

| | Deposit = TOP: MEDIUM (26.0/11.0)

| BRANCH_NAME = DESSIE BRANCH

| | Deposit = BASIC: SMALL (24.0/6.0)

| | Deposit = MSME: MEDIUM (5.0/2.0)

| | Deposit = TOP: MEDIUM (9.0/3.0)

| BRANCH_NAME = GONDAR BRANCH: SMALL (76.0/17.0)

| BRANCH_NAME = BEKLO BET BRANCH

| | LIMIT_FACILITY = NO

| | | Deposit = BASIC: SMALL (130.0/49.0)

| | | Deposit = MSME: SMALL (21.0/10.0)

| | | Deposit = TOP

| | | | CCY = ETB: MEDIUM (71.0/40.0)

| | | | CCY = USD: SMALL (4.0/2.0)

| | | | CCY = EUR: MEDIUM (0.0)

| | | | CCY = GBP: MEDIUM (0.0)

| | LIMIT_FACILITY = YES: CORPORATE (30.0/17.0)

| BRANCH_NAME = BOLE MEDHANIALEM BRANCH

| | LIMIT_FACILITY = NO

| | | Deposit = BASIC: SMALL (91.0/36.0)

| | | Deposit = MSME: SMALL (27.0/14.0)

| | | Deposit = TOP: MEDIUM (42.0/21.0)

| | LIMIT_FACILITY = YES: MEDIUM (12.0/4.0)

| BRANCH_NAME = AMOUDI BRANCH

| | CCY = ETB

| | | Deposit = BASIC: CORPORATE (143.0/89.0)

| | | Deposit = MSME: SMALL (12.0/5.0)

| | | Deposit = TOP: CORPORATE (59.0/34.0)

| | CCY = USD

| | | Deposit = BASIC: CORPORATE (82.0/49.0)

| | | Deposit = MSME: SMALL (4.0/2.0)

| | | Deposit = TOP: CORPORATE (31.0/18.0)

| | CCY = EUR: SMALL (31.0/8.0)

| | CCY = GBP: SMALL (5.0/1.0)

| BRANCH_NAME = SARIS BRANCH

| | LIMIT_FACILITY = NO

| | | Deposit = BASIC: SMALL (48.0/13.0)

| | | Deposit = MSME: MEDIUM (12.0/6.0)

| | | Deposit = TOP: SMALL (18.0/8.0)

| | LIMIT_FACILITY = YES: MEDIUM (12.0/6.0)

| BRANCH_NAME = DIRE DAWA BRANCH

| | Deposit = BASIC: SMALL (58.0/16.0)

| | Deposit = MSME: SMALL (19.0/2.0)

| | Deposit = TOP: MEDIUM (25.0/11.0)

| BRANCH_NAME = KERRA BRANCH

| | LIMIT_FACILITY = NO: SMALL (171.0/34.0)

| | LIMIT_FACILITY = YES: MEDIUM (8.0/4.0)

| BRANCH_NAME = WELLO SEFER PREMIER BRANCH

| | CCY = ETB

| | | LIMIT_FACILITY = NO

| | | | Deposit = BASIC: SMALL (227.0/137.0)

| | | | Deposit = MSME: SMALL (37.0/20.0)

| | | | Deposit = TOP: CORPORATE (202.0/125.0)

| | | LIMIT_FACILITY = YES: MEDIUM (2.0/1.0)

| | CCY = USD

| | | Deposit = BASIC: CORPORATE (20.0/7.0)

| | | Deposit = MSME: MEDIUM (1.0)

| | | Deposit = TOP: CORPORATE (18.0/5.0)

| | CCY = EUR: SMALL (4.0/1.0)

| | CCY = GBP: SMALL (4.0/2.0)

| BRANCH_NAME = MEKELE BRANCH: CORPORATE (5.0)

| BRANCH_NAME = TANA BRANCH: CORPORATE (5.0)

| BRANCH_NAME = ADIGRAT BRANCH: CORPORATE (1.0)

Number of Leaves : 85

Size of the tree : 117

Time taken to build model: 0.66 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.03 seconds

=== Summary ===

Correctly Classified Instances 12624 92.0855 %

Incorrectly Classified Instances 1085 7.9145 %

Kappa statistic 0.2617

Mean absolute error 0.0824

Root mean squared error 0.2045

Relative absolute error 79.6367 %

Root relative squared error 89.892 %

Total Number of Instances 13709

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.991	0.761	0.935	0.991	0.962	0.387	0.807	0.968	SMALL
	0.115	0.009	0.420	0.115	0.181	0.199	0.714	0.203	MEDIUM
	0.183	0.008	0.420	0.183	0.255	0.262	0.925	0.313	CORPORATE
Weighted Avg.	0.921	0.698	0.892	0.921	0.900	0.373	0.806	0.908	

=== Confusion Matrix ===

a	b	c	<-- classified as
12464	54	55	a = SMALL
569	81	54	b = MEDIUM
295	58	79	c = CORPORATE